

VOL. 1.

Charles Platt

Encyclopedia of Electronic Components



Power Sources & Conversion

Resistors • Capacitors • Inductors
Switches • Encoders • Relays • Transistors



O'REILLY®

Make:
makezine.com

Encyclopedia of Electronic Components Volume 1

Charles Platt

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

Encyclopedia of Electronic Components Volume 1

by Charles Platt

Copyright © 2013 Helpful Corporation. All rights reserved.

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (<http://my.safaribooksonline.com>). For more information, contact our corporate/institutional sales department: 800-998-9938 or corporate@oreilly.com.

Editor: Brian Jepson

Production Editor: Melanie Yarbrough

Proofreader: Melanie Yarbrough

Indexer: Judy McConville

Cover Designer: Mark Paglietti

Interior Designer: Edie Freedman and Nellie McKesson

Illustrator: Charles Platt

Photographer: Charles Platt

Cover Production: Randy Comer

October 2012: First Edition

Revision History for the First Edition:

2012-10-03 First release

See <http://oreilly.com/catalog/errata.csp?isbn=9781449333898> for release details.

Nutshell Handbook, the Nutshell Handbook logo, and the O'Reilly logo are registered trademarks of O'Reilly Media, Inc. [Encyclopedia of Electronic Components Volume 1](#), the cover images, and related trade dress are trademarks of O'Reilly Media, Inc.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and O'Reilly Media, Inc., was aware of a trademark claim, the designations have been printed in caps or initial caps.

While every precaution has been taken in the preparation of this book, the publisher and authors assume no responsibility for errors or omissions, or for damages resulting from the use of the information contained herein.

ISBN: 978-1-449-33389-8

[TI]

To Mark Frauenfelder, who reacquainted me with the pleasures of Making.

Table of Contents

Preface	xix
----------------------	------------

1. How to Use This Book	1
Reference vs. Tutorial	1
Theory and Practice	1
Organization	1
Subject Paths	2
Inclusions and Exclusions	2
Typographical Conventions	3
Volume Contents	3
Safari® Books Online	3
How to Contact Us	4

> POWER

> > SOURCE

2. Battery	5
What It Does	5
How It Works	6
Electrode Terminology	7
Variants	7
Disposable Batteries	8
Rechargeable Batteries	9
Values	11
Amperage	11
Capacity	11
Voltage	13
How To Use It	14

What Can Go Wrong	15
Short Circuits: Overheating And Fire	15
Diminished Performance Caused By Improper Recharging ..	15
Complete Discharge Of Lead-Acid Battery	15
Inadequate Current	15
Incorrect Polarity	15
Reverse Charging	16
Sulfurization	16
High Current Flow Between Parallel Batteries	16

> > CONNECTION

3. Jumper	17
What It Does	17
How It Works	17
Variants	18
Values	18
How To Use It	19
What Can Go Wrong	19
 4. Fuse	 21
What It Does	21
How It Works	21
Values	22
Variants	22
Small Cartridge Fuses	23
Automotive Fuses	23
Strip Fuses	24
Through-Hole Fuses	24
Resettable Fuses	24
Surface Mount Fuses	26
How To Use It	26
What Can Go Wrong	27
Repeated Failure	27
Soldering Damage	27
Placement	28
 5. Pushbutton	 29
What It Does	29
How It Works	29
Variants	30
Poles And Throws	30
On-Off Behavior	30
Slider	31
Styles	31
Termination And Contact Plating	32
Mounting Style	32
Sealed Or Unsealed	32

Latching	33
Foot Pedal	33
Keypad	33
Tactile Switch	34
Membrane Pad	34
Radio Buttons	35
Snap-Action Switches	35
Emergency Switch	35
Values	35
How To Use It	35
What Can Go Wrong	35
No Button	35
Mounting Problems	35
LED Issues	36
Other Problems	36
6. Switch	37
What It Does	37
How It Works	37
Variants	38
Terminology	38
Poles And Throws	38
On-Off Behavior	39
Snap-Action	39
Rocker	40
Slider	40
Toggle	41
DIP	43
SIP	44
Paddle	44
Vandal Resistant Switch	45
Tactile Switch	45
Mounting Options	45
Termination	45
Contact Plating Options	45
Values	45
How To Use It	46
Power Switches	46
Limit Switches	46
Logic Circuits	47
Alternatives	47
What Can Go Wrong	47
Arcing	47
Dry Joints	48
Short Circuits	48
Contact Contamination	48
Wrong Terminal Type	48
Contact Bounce	48

Mechanical Wear	48
Mounting Problems	48
Cryptic Schematics	49
7. Rotary Switch	51
What It Does	51
How It Works	52
Variants	52
Conventional	52
Rotary DIP	53
Gray Code	54
PC Board Rotary Switch	55
Mechanical Encoder	55
Pushwheel And Thumbwheel	55
Keylock	55
Values	56
How To Use It	56
What Can Go Wrong	57
Vulnerable Contacts	57
Contact Overload	57
Misalignment	57
Misidentified Shorting Switch	57
User Abuse	57
Wrong Shaft, Wrong Knobs, Nuts That Get Lost, Too Big To Fit	57
8. Rotational Encoder	59
What It Does	59
How It Works	59
Variants	60
Pulses And Detents	61
Format	61
Output	61
Rotational Resistance	61
Values	61
Contact Bounce	61
Sliding Noise	62
How To Use It	62
What Can Go Wrong	62
Switch Bounce	62
Contact Burnout	63
9. Relay	65
What It Does	65
How It Works	66
Variants	67
Latching	67
Polarity	67
Pinout Variations	67

Reed Relay	68
Small Signal Relay	68
Automotive Relays	69
General Purpose/Industrial	69
Time Delay Relay	69
Contactor	70
Values	70
How To Use It	71
What Can Go Wrong	72
Wrong Pinouts	72
Wrong Orientation	72
Wrong Type	72
Wrong Polarity	72
AC And DC	72
Chatter	72
Relay Coil Voltage Spike	72
Arcing	72
Magnetic Fields	72
Environmental Hazards	73

> > MODERATION

10. Resistor	75
What It Does	75
How It Works	76
Variants	76
Resistor Array	77
Values	79
Tolerance	79
Value Coding	81
Stability	82
Materials	82
How To Use It	84
In Series With LED	84
Current Limiting With A Transistor	84
Pullup And Pulldown Resistors	85
Audio Tone Control	85
RC Network	85
Voltage Divider	86
Resistors In Series	86
Resistors In Parallel	86
What Can Go Wrong	87
Heat	87
Noise	87
Inductance	87
Inaccuracy	87

Wrong Values	88
11. Potentiometer	89
What It Does	89
How It Works	90
Variants	90
Linear And Log Taper	90
Classic-Style Potentiometer	91
Multiple-Turn Potentiometer	92
Ganged Potentiometer	93
Switched Potentiometer	93
Slider Potentiometer	93
Trimmer Potentiometer	93
How To Use It	94
What Can Go Wrong	95
Wear And Tear	95
Knobs That Don't Fit	95
Nuts That Get Lost	95
A Shaft That Isn't Long Enough	96
Sliders With No Finger Grip	96
Too Big To Fit	96
Overheating	96
The Wrong Taper	96
12. Capacitor	97
What It Does	97
How It Works	97
Variants	99
Format	99
Principal Types	101
Dielectrics	103
Values	104
Farads	104
Commonly Used Values	104
Dielectric Constant	105
The Time Constant	105
Multiple Capacitors	106
Alternating Current And Capacitive Reactance	106
Equivalent Series Resistance	106
How To Use It	107
Bypass Capacitor	107
Coupling Capacitor	107
High-Pass Filter	107
Low-Pass Filter	107
Smoothing Capacitor	108
Snubber	108
Capacitor As A Battery Substitute	109
What Can Go Wrong	109

Wrong Polarity	110
Voltage Overload	110
Leakage	110
Dielectric Memory	110
Specific Electrolytic Issues	110
Heat	110
Vibration	110
Misleading Nomenclature	111
13. Variable Capacitor	113
What It Does	113
How It Works	113
Variants	114
Values	115
Formats	115
How To Use It	115
What Can Go Wrong	117
Failure To Ground Trimmer Capacitor While Adjusting It ...	117
Application Of Overcoat Material Or “Lock Paint”	117
Lack Of Shielding	117
14. Inductor	119
What It Does	119
How It Works	120
DC Through A Coil	121
Magnetic Core	122
EMF And Back-EMF	122
Electrical And Magnetic Polarity	123
Variants	124
Magnetic Cores	124
Nonmagnetic Cores	125
Variable Inductors	125
Ferrite Beads	126
Toroidal Cores	126
Gyrator	127
Values	128
Calculating Inductance	128
Calculating Reactance	128
Calculating Reluctance	129
Datasheet Terminology	129
Series And Parallel Configurations	129
Time Constant	129
How To Use It	130
Core Choices	132
Miniaturization	132
What Can Go Wrong	132
Real-World Defects	132
Saturation	132

RF Problems	133
-------------------	-----

> > CONVERSION

15. AC-AC Transformer 135

What It Does	135
How It Works	136
The Core	137
Taps	137
Variants	138
Core Shapes	138
Power Transformer	138
Plug-In Transformer	139
Isolation Transformer	139
Autotransformer	140
Variable Transformer	140
Audio Transformer	140
Split-Bobbin Transformer	141
Surface-Mount Transformer	141
Values	141
How To Use It	142
What Can Go Wrong	142
Reversal Of Input And Output	142
Shock Hazard From Common Ground	142
Accidental DC Input	142
Overload	142
Incorrect AC Frequency	142

16. AC-DC Power Supply 143

What It Does	143
Variants	143
Linear Regulated Power Supply	143
Switching Power Supply	144
Unregulated Power Supply	146
Adjustable Power Supply	146
Voltage Multiplier	146
Formats	146
How To Use It	147
What Can Go Wrong	147
High Voltage Shock	147
Capacitor Failure	147
Electrical Noise	147
Peak Inrush	147

17. DC-DC Converter 149

What It Does	149
How It Works	149
Variants	150

Buck Converter	150
Boost Converter	151
Flyback Converter With Inductor	151
Flyback Converter With Transformer	151
Formats	151
Values	152
Nominal Input Voltage And Frequency	152
Output Voltage	153
Input Current And Output Current	153
Load Regulation	153
Efficiency	153
Ripple And Noise	154
Isolated Or Non-Isolated	154
How To Use It	154
What Can Go Wrong	155
Electrical Noise In Output	155
Excess Heat With No Load	155
Inaccurate Voltage Output With Low Load	155
18. DC-AC Inverter	157
What It Does	157
How It Works	157
Variants	158
Values	158
How To Use It	159
What Can Go Wrong	160
> > REGULATION	
19. Voltage Regulator	161
What It Does	161
How It Works	161
Variants	163
Packaging	163
Popular Varieties	163
Adjustable Regulators	163
Negative And Positive Regulators	164
Low-Dropout Linear Regulators	164
Quasi-Low-Dropout Linear Regulators	165
Additional Pin Functions	165
Values	165
How To Use It	165
What Can Go Wrong	166
Inadequate Heat Management	166
Transient Response	166
Misidentified Parts	166
Misidentified Pins	167
Dropout Caused By Low Battery	167

Inaccurate Delivered Voltage	167
------------------------------------	-----

> ELECTROMAGNETISM

> > LINEAR

20. Electromagnet	169
What It Does	169
How It Works	169
Variants	170
Values	171
How To Use It	171
What Can Go Wrong	172
 21. Solenoid	 173
What It Does	173
How It Works	174
Variants	176
Low Profile	176
Latching	176
Rotary	176
Hinged Clapper	176
Values	176
Coil Size Vs. Power	177
How To Use It	177
What Can Go Wrong	177
Heat	177
AC Inrush	177
Unwanted EMF	177
Loose Plunger	177

> > ROTATIONAL

22. DC Motor	179
What It Does	179
How It Works	179
Variants	181
Coil Configurations	181
Gearhead Motor	181
Brushless DC Motor	183
Linear Actuator	184
Values	184
How To Use It	185
Speed Control	186
Direction Control	186
Limit Switches	187
What Can Go Wrong	187

Brushes And Commutator	187
Electrical Noise	187
Heat Effects	188
Ambient Conditions	188
Wrong Shaft Type Or Diameter	188
Incompatible Motor Mounts	188
Backlash	188
Bearings	188
Audible Noise	189
23. AC Motor	191
What It Does	191
How It Works	191
Stator Design	191
Rotor Design	192
Variants	195
Single-Phase Induction Motor	195
Three-Phase Induction Motor	196
Synchronous Motor	196
Reluctance Motor	197
Variable Frequency Drive	198
Wound-Rotor AC Induction Motor	198
Universal Motor	198
Inverted AC Motors	199
Values	199
How To Use It	199
What Can Go Wrong	200
Premature Restart	200
Frequent Restart	200
Undervoltage Or Voltage Imbalance	200
Stalled Motor	200
Protective Relays	200
Excess Torque	200
Internal Breakage	200
24. Servo Motor	201
What It Does	201
How It Works	201
Variants	203
Values	204
How To Use It	205
Modification For Continuous Rotation	206
What Can Go Wrong	206
Incorrect Wiring	206
Shaft/Horn Mismatch	206
Unrealistically Rapid Software Commands	207
Jitter	207
Motor Overload	207

Unrealistic Duty Cycle	207
Electrical Noise	207
25. Stepper Motor	209
What It Does	209
How It Works	209
Reluctance Stepper Motors	210
Permanent Magnet Stepper Motors	211
Bipolar Stepper Motors	213
Unipolar Motors	213
Variants	214
High Phase Count	214
Hybrid	216
Bifilar	216
Multiphase	216
Microstepping	217
Sensing And Feedback	217
Voltage Control	217
Values	218
How To Use It	218
Protection Diodes	218
Positional Control	219
What Can Go Wrong	219
Incorrect Wiring	219
Step Loss	219
Excessive Torque	219
Hysteresis	220
Resonance	220
Hunting	220
Saturation	220
Rotor Demagnetization	220

> DISCRETE SEMICONDUCTOR

> > SINGLE JUNCTION

26. Diode	221
What It Does	221
How It Works	223
Variants	224
Packaging	224
Signal Diodes	224
Rectifier Diodes	224
Zener Diode	224
Transient Voltage Suppressor (TVS)	225
Schottky Diode	225
Varactor Diode	225

Tunnel Diode, Gunn Diode, PIN Diode	226
Diode Array	226
Bridge Rectifier	226
Values	226
How To Use It	227
Rectification	227
Back-EMF Suppression	228
Voltage Selection	229
Voltage Clamping	230
Logic Gate	230
DC Voltage Regulation And Noise Suppression	230
AC Voltage Control And Signal Clipping	231
Voltage Sensing	231
What Can Go Wrong	232
Overload	232
Reversed Polarity	233
Wrong Type Of Diode	233
27. Unijunction Transistor	235
What It Does	235
How It Works	236
Variants	238
Values	238
How To Use It	239
What Can Go Wrong	239
Name Confusion	239
Incorrect Bias	239
Overload	240
 > > MULTI-JUNCTION	
28. Bipolar Transistor	241
What It Does	241
How It Works	241
Current Gain	244
Terminology	245
Variants	245
Packaging	245
Connections	246
How To Use It	246
Darlington Pairs	248
Amplifiers	250
What Can Go Wrong	251
Wrong Connections On A Bipolar Transistor	251
Wrong Connections On A Darlington Pair Chip	251
Soldering Damage	252
Excessive Current Or Voltage	252

Excessive Leakage	252
29. Field Effect Transistor	253
What It Does	253
How It Works	253
JFETs	253
JFET Behavior	255
MOSFETs	256
The Substrate Connection	261
Variants	262
MESFET	262
V-Channel MOSFET	262
Trench MOS	262
Values	262
How To Use It	263
P-Channel Disadvantage	263
Bipolar Substitution	263
Amplifier Front Ends	263
Voltage-Controlled Resistor	263
Compatibility With Digital Devices	263
What Can Go Wrong	263
Static Electricity	263
Heat	263
Wrong Bias	264
Appendix A. Schematic Symbols	265
Index	269

Preface

At a time when information is widely and freely available in greater quantities than ever before, the reader may wonder whether *The Encyclopedia of Electronic Components* is really necessary. Surely, anything you want to know can be found online?

Well, yes and no. Let's consider the available resources.

1. Datasheets

Datasheets are indispensable, but they have limitations. Some are detailed; others are skimpy. Some show you sample schematics as a guide to using a component; many don't. None of them tells you much about how a component works, because that's not their purpose. Often they don't mention other components that must be added. Some datasheets for DC-DC converters, for instance, say nothing at all about bypass capacitors, even though the capacitors may be essential. A datasheet for an optocoupler says nothing about the pullup resistor required by the open-collector output.

Datasheets don't facilitate comparison shopping. A datasheet from one manufacturer will not compare its products with those from another manufacturer, and may not even provide much guidance about alternatives that are available from the same manufacturer. For example, a da-

tasheet for a linear voltage regulator won't suggest that you might do better to use a DC-DC converter in an application where high efficiency is important.

Most of all, datasheets don't tell you how to avoid common mistakes. What actually happens if you connect that tantalum capacitor the wrong way around? A datasheet gives you the customary list of absolute maximum values, and after that, you are on your own, burning things out, encountering mysterious electronic behavior, and discovering limitations that are so well known, the datasheet didn't bother to mention them. In my experience, relying on datasheets creates a significant risk of reinventing the wheel.

2. Wikipedia

Wikipedia's coverage of electronics is impressive but inconsistent. Some entries are elementary, while others are extremely technical. Some are shallow, while others are deep. Some are well organized, while others run off into obscure topics that may have interested one of the contributors but are of little practical value to most readers. Many topics are distributed over multiple entries, forcing you to hunt through several URLs. Overall, Wikipedia tends to be good if you want theory, but not-so-good if you want hands-on practicality.

3. Manufacturers' Tutorials

A few helpful and enlightened manufacturers have compiled highly authoritative, instructional overviews of the components that they sell. Lit-telfuse, for instance, publishes an excellent series of documents telling you everything you could possibly want to know about fuses. But now you encounter a different problem: There is so much information, you'll need a couple of hours to dig through it all. Also, because the tutorials tend not to receive high page rankings on Google, they can be hard to find. And if a manufacturer has gaps in its product line, its tutorial is unlikely to mention them. Consequently, you won't know what's missing.

4. Personal Guides

It is a well-known attribute of the Web that many individuals feel the impulse to share everything they know (or think they know) about a particular topic. These personal guides can present surprisingly thorough online coverage of relatively obscure issues, such as the types of capacitors most suitable for loudspeaker crossover circuits, or the correct derivation of amp-hour numbers for lead-acid batteries. Unfortunately, on some sites you can also find errors, unsubstantiated opinions, plagiarism, and eccentricity. My general rule is that three or more guides generally have to agree with each other before their statements can be trusted—and even then, I have a small residue of doubt. The search-inspect-and-verify process can take a while.

So—yes, the information that you want usually does exist somewhere online, but no, it may not be easy to find. The vastness of the Web is not organized like an encyclopedia.

What about books? Generally speaking, they tend to be entry-level, or they specialize in narrow areas. A few broad-ranging books are truly excellent, but they are primarily educational, organized in an instructional sequence. They are not reference books.

The Encyclopedic Solution

Scarcity or inaccessibility of information ceased to be a problem many years ago. Its vast quantity, inconsistency, and dispersal have become the new barriers to acquiring knowledge. If you have to go hunting among datasheets, Wikipedia, manufacturers' tutorials (which may or may not exist), personal guides (which may have unrevealed bias), and multiple educational books, the process will be inconvenient and time-consuming. If you plan to revisit the topic in the future, you'll have to remember which URLs were useful and which ones weren't—and you may find that many of them are not even there anymore.

When I considered these issues during my own work as an electronics columnist for *Make* magazine, I saw a real need for a fact-checked, cross-referenced encyclopedia that would compile the basic information about components concisely, in an organized, consistent format, with informative photographs, schematics, and diagrams. It might save many people a lot of search time if it could summarize how components work, how to use them, what the alternatives are, and what the common errors and problems may be.

That is the modest ambition of *The Encyclopedia of Electronic Components*.

The Audience

Like any reference work, this one hopes to serve two categories of readers: The informed and the not-yet-informed.

Perhaps you are learning electronics, and you see a part listed in a catalog. It looks interesting, but the catalog doesn't tell you exactly what the part does or how it is commonly used. You need to look it up either by function or by name, but you're not sure where to start. An encyclopedic reference can simplify the fact-finding process, can save you from ordering a part that may be inappropriate, and can tell you how it should be used.

Perhaps, instead, you are an electronics engineer or hobbyist, thinking about a new circuit. You remember using a component three or four years ago, but your recollection may not be reliable. You need to refresh your memory with a quick summary—and so, you open the encyclopedia, just to make sure.

Completeness

Obviously, this book cannot include every component that exists. Mouser Electronics claims to have more than 2 million products listed in its online database. *The Encyclopedia of Electronic Components* only has room for a fraction of that number—but still, it can refer you to the primary types. The electronic edition of this book should allow easy insertions and updates. My hope is that it can become an ever-expanding resource.

Acknowledgments

Any reference work draws inspiration from many sources, and this one is no exception. Three were of special importance:

Practical Electronics for Inventors by Paul Scherz (second edition) McGraw-Hill, 2007

Electronic Devices and Circuit Theory by Robert L. Boylestad and Louis Nashelsky (ninth edition) Pearson Education Inc., 2006

The Art of Electronics by Paul Horowitz and Winfield Hill (second edition) Cambridge University Press, 2006

I also made extensive use of information gleaned through Mouser Electronics and Jameco Electronics. And where would any of us be without *Getting Started in Electronics* by Forrest M. Mims III, or *The TTL Cookbook* by Don Lancaster?

In addition, there were individuals who provided special assistance. My editor, Brian Jepson, was immensely helpful in the development of the project. Michael Butler contributed greatly to the early concept and its structure. Josh Gates did resourceful research. My publishers, O'Reilly Media, demonstrated their faith in my work. Kevin Kelly unwittingly influenced me with his legendary interest in “access to tools.”

Primary fact checkers were Eric Moberg, Chris Lirakis, Jason George, Roy Rabey, Emre Tuncer, and Patrick Fagg. I am indebted to them for their help. Any remaining errors are, of course, my responsibility.

Lastly I should mention my school friends from decades ago: Hugh Levinson, Patrick Fagg, Graham Rogers, William Edmondson, and John Witty, who helped me to feel that it was okay to be a nerdy kid building my own audio equipment, long before the word “nerd” existed.

—Charles Platt, 2012

How to Use This Book

1

To avoid misunderstandings regarding the purpose and method of this book, here is a quick guide regarding the way in which it has been conceived and organized.

Reference vs. Tutorial

As its title suggests, this is a reference book, not a tutorial. In other words, it does not begin with elementary concepts and build sequentially toward concepts that are more advanced.

You should be able to dip into the text at any point, locate the topic that interests you, learn what you need to know, and then put the book aside. If you choose to read it straight through from beginning to end, you will not find concepts being introduced in a sequential, cumulative manner.

My book *Make:Electronics* follows the tutorial approach. Its range, however, is more circumscribed than that of this encyclopedia, because a tutorial inevitably allocates a lot of space to step-by-step explanations and instructions.

Theory and Practice

This book is oriented toward practicality rather than theory. I am assuming that the reader mostly wants to know how to use electronic components, rather than why they work the way they

do. Consequently I have not included any proofs of formulae, any definitions rooted in electrical theory, or any historical background. Units are defined only to the extent that is necessary to avoid confusion.

Many books on electronics theory already exist, if theory is of interest to you.

Organization

The encyclopedia is divided into entries, each entry being devoted to one broad type of component. Two rules determine whether a component has an entry all to itself, or is subsumed into another entry:

1. A component merits its own entry if it is (a) widely used or (b) not-so-widely used but has a unique identity and maybe some historical status. A widely used component would be a **bipolar transistor**, while a not-so-widely-used component with a unique identity would be a **unijunction transistor**.
2. A component does not merit its own entry if it is (a) seldom used or (b) very similar in function to another component that is more widely used. For example, the *rheostat* is sub-

sumed into the **potentiometer** section, while *silicon diode*, *Zener diode*, and *germanium diode* are combined together in the **diode** entry.

Inevitably, these guidelines required judgment calls that in some cases may seem arbitrary. My ultimate decision was based on where I would expect to find a component if I was looking for it myself.

Subject Paths

Entries are not organized alphabetically. Instead they are grouped by subject, in much the same way that books in the nonfiction section of a library are organized by the Dewey Decimal System. This is convenient if you don't know exactly what you are looking for, or if you don't know all the options that may be available to perform a task that you have in mind.

Each primary category is divided into subcategories, and the subcategories are divided into component types. This hierarchy is shown in [Figure 1-1](#). It is also apparent when you look at the top of the first page of each entry, where you will find the path that leads to it. The **capacitor** entry, for instance, is headed with this path:

power > moderation > capacitor

Any classification scheme tends to run into exceptions. You can buy a chip containing a *resistor array*, for instance. Technically, this is an *analog integrated circuit*, but should it really be included with solid-state relays and comparators? A decision was made to put it in the **resistor** section, because this seemed more useful.

Some components have hybrid functions. In Volume 2, in the *integrated circuit* subcategory, we will distinguish between those that are *analog* and those that are *digital*. So where should an **analog-digital converter** be listed? It will be found under *analog*, because that category seems better associated with its primary function, and people may be more likely to look for it there.

Primary Category	Secondary Category	Component Type
power	source	battery
		jumper
		fuse
		pushbutton
		switch
		rotary switch
		rotational encoder
	moderation	relay
		resistor
		potentiometer
		capacitor
		variable capacitor
	conversion	inductor
		AC-AC transformer
		AC-DC power supply
		DC-DC converter
		DC-AC inverter
	regulation	voltage regulator
electro-magnetism	linear output	electromagnet
		solenoid
	rotational output	DC motor
		AC motor
		servo motor
discrete semi-conductor	single junction	diode
		unijunction transistor
	multi-junction	bipolar transistor
		field-effect transistor

Figure 1-1. The subject-oriented organization of categories and entries in this encyclopedia.

Inclusions and Exclusions

There is also the question of what is, and what is not, a component. Is wire a component? Not for

the purposes of this encyclopedia. How about a **DC-DC converter**? Because converters are now sold in small packages by component suppliers, they have been included as components.

Many similar decisions had to be made on a case-by-case basis. Undoubtedly, some readers will disagree with the outcome, but reconciling all the disagreements would have been impossible. Speaking personally, the best I could do was create a book that is organized in the way that would suit me best if I were using it myself.

Typographical Conventions

Throughout this encyclopedia, the names of components that have their own entries are presented in **bold type**. Other important electronics terms or component names are presented in *italics* where they first appear in any one section.

The names of components, and the categories to which they belong, are all set in lower-case type, except where a term is normally capitalized because it is an acronym or a trademark. *Trimpot*, for instance, is trademarked by Bourns, but *trimmer* is not. **LED** is an acronym, but *cap* (abbreviation for **capacitor**) is not.

Where formulae are used, they are expressed in a format that will be familiar to computer programmers but may be unfamiliar to others. The * (asterisk) symbol is used in place of a multiplication sign, while the / (slash symbol) is used to indicate division. Where pairs of parentheses are nested, the most deeply nested pair identifies the operations that should be performed first.

Volume Contents

Practical considerations relating to book length influenced the decision to divide *The Encyclopedia of Electronic Components* into three volumes. Each volume deals with broad subject areas as follows.

Volume 1

Power, electromagnetism, and discrete semiconductors.

The *power* category includes sources of power and methods to distribute, store, interrupt, and modify power. The *electromagnetism* category includes devices that exert force linearly, and others that create a turning force. *Discrete semiconductors* include the main types of diodes and transistors.

Volume 2

Integrated circuits, light sources, sound sources, heat sources, and high-frequency sources.

Integrated circuits are divided into analog and digital components. *Light sources* range from incandescent bulbs to LEDs and small display screens; some reflective components, such as liquid-crystal displays and e-ink, are also included. *Sound sources* are primarily electromagnetic.

Volume 3

Sensing devices.

The field of sensors has become so extensive, they easily merit a volume to themselves. *Sensing devices* include those that detect light, sound, heat, motion, pressure, gas, humidity, orientation, electricity, proximity, force, and radiation.

At the time of writing, volumes 2 and 3 are still in preparation, but their contents are expected to be as described above.

Safari® Books Online

Safari Books Online is an on-demand digital library that lets you easily search over 7,500 technology and creative reference books and videos to find the answers you need quickly.

With a subscription, you can read any page and watch any video from our library online. Read books on your cell phone and mobile devices. Access new titles before they are available for print, and get exclusive access to manuscripts in development and post feedback for the authors.

Copy and paste code samples, organize your favorites, download chapters, bookmark key sections, create notes, print out pages, and benefit from tons of other time-saving features.

O'Reilly Media has uploaded this book to the Safari Books Online service. To have full digital access to this book and others on similar topics from O'Reilly and other publishers, sign up for free at <http://my.safaribooksonline.com>.

How to Contact Us

Please address comments and questions concerning this book to the publisher:

MAKE
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (in the United States or
Canada)
707-829-0515 (international or local)
707-829-0104 (fax)

MAKE unites, inspires, informs, and entertains a growing community of resourceful people who undertake amazing projects in their backyards, basements, and garages. MAKE celebrates your right to tweak, hack, and bend any technology to your will. The MAKE audience continues to be a growing culture and community that believes in bettering ourselves, our environment, our ed-

ucational system—our entire world. This is much more than an audience, it's a worldwide movement that Make is leading—we call it the Maker Movement.

For more information about MAKE, visit us online:

MAKE magazine: <http://makezine.com/magazine/>

Maker Faire: <http://makerfaire.com>

Makezine.com: <http://makezine.com>

Maker Shed: <http://makershed.com/>

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at:

http://oreil.ly/encyc_electronic_comp_v1

To comment or ask technical questions about this book, send email to:

bookquestions@oreilly.com

For more information about our books, courses, conferences, and news, see our website at <http://www.oreilly.com>.

Find us on Facebook: <http://facebook.com/oreilly>

Follow us on Twitter: <http://twitter.com/oreillymedia>

Watch us on YouTube: <http://www.youtube.com/oreillymedia>

battery

2

This entry covers electrochemical power sources. Electricity is most often generated electromagnetically, but since these sources cannot be classified as components, they are outside the scope of the encyclopedia. Electrostatic sources are excluded for similar reasons.

A battery is sometimes referred to as a *cell* or *power cell*, but can actually contain multiple cells, as defined in this entry. It used to be called an *accumulator* or a *pile*, but those terms are now archaic.

OTHER RELATED COMPONENTS

- **capacitor** (see [Chapter 12](#))

What It Does

A battery contains one or more *electrochemical cells* in which chemical reactions create an electrical potential between two immersed terminals. This potential can be discharged as *current* passing through a *load*.

An electrochemical cell should not be confused with an *electrolytic cell*, which is powered by an external source of electricity to promote *electrolysis*, whereby chemical compounds are broken down to their constituent elements. An electrolytic cell thus consumes electricity, while an electrochemical cell produces electricity.

Batteries range in size from *button cells* to large *lead-acid* units that store power generated by solar panels or windmills in locations that can be off the grid. Arrays of large batteries can provide bridging power for businesses or even small communities where conventional power is unreliable. [Figure 2-1](#) shows a 60KW, 480VDC self-watering battery array installed in a corporate data center, supplementing wind and solar sources

and providing time-of-day peak shaving of energy usage. Each lead-acid battery in this array measures approximately 28" × 24" × 12" and weighs about 1,000 lb.



Figure 2-1. A battery array providing 60KW at 480VDC as backup for a corporate data center. (Photo by permission of Hybridyne Power Systems, Canada, Inc., and the Hybridyne group of companies. Copyright by Hybridyne, an internationally registered trademark of Hybridyne Power Systems Canada Inc. No right of further reproduction unless specifically granted by Hybridyne.)

Schematic symbols for a battery are shown in [Figure 2-2](#). The longer of the two lines represents the positive side of the battery, in each case. One way to remember this is by imagining that the longer line can be snipped in half so that the two segments can combine to form a + sign. Traditionally, multiple connected battery symbols indicate multiple cells inside a battery; thus the center symbols in the figure could indicate a 3V battery, while those on the right would indicate a voltage greater than 3V. In practice, this convention is not followed conscientiously.

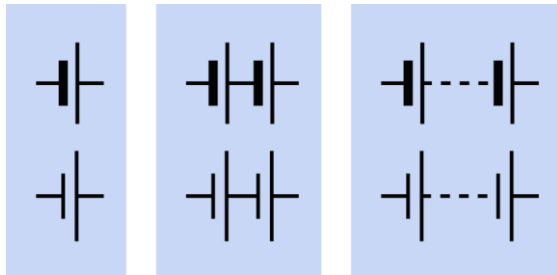


Figure 2-2. Schematic symbols for a battery. Each pair of symbols within a blue rectangle is functionally identical.

How It Works

In a basic battery design often used for demonstration purposes, a piece of copper serves as an *electrode*, partially immersed in a solution of copper sulfate, while a piece of zinc forms a second electrode, partially immersed in a solution of zinc sulfate. Each sulfate solution is known as an *electrolyte*, the complete battery may be referred to as a *cell*, and each half of it may be termed a *half-cell*.

A simplified cross-section view is shown in [Figure 2-3](#). Blue arrows show the movement of electrons from the zinc terminal (the *anode*), through an external load, and into a copper terminal (the *cathode*). A *membrane separator* allows the electrons to circulate back through the battery, while preventing electrolyte mixing.

Orange arrows represent positive copper *ions*. White arrows represent positive zinc ions. (An ion

is an atom with an excess or deficit of electrons.) The zinc ions are attracted into the zinc sulfate electrolyte, resulting in a net loss of mass from the zinc electrode.

Meanwhile, electrons passing into the copper electrode tend to attract positive copper ions, shown as orange arrows in the diagram. The copper ions are drawn out of the copper sulfate electrolyte, and result in a net accumulation of copper atoms on the copper electrode.

This process is energized partially by the fact that zinc tends to lose electrons more easily than copper.

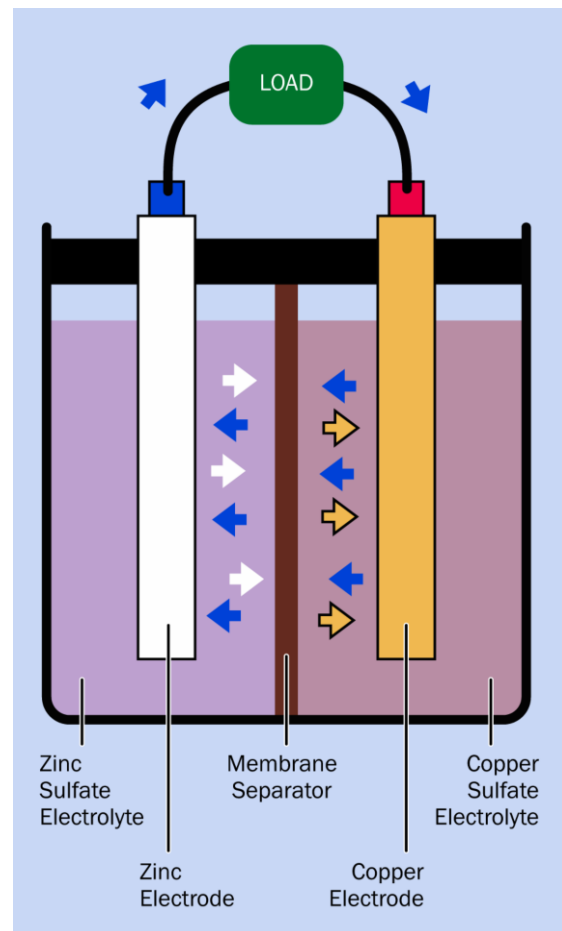


Figure 2-3. A classically simple electrochemical cell. See text for additional details.

Batteries for use in consumer electronics typically use a paste instead of a liquid as an electrolyte, and have been referred to as *dry cells*, although this term is becoming obsolete. The two half-cells may be combined concentrically, as in a typical 1.5-volt C, D, AA, or AAA alkaline battery (see Figure 2-4).

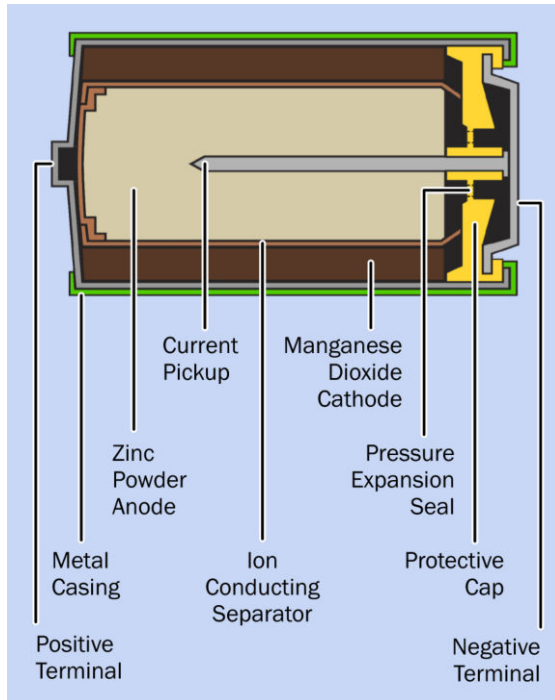


Figure 2-4. Cross-section view of a typical 1.5-volt alkaline battery.

A 1.5V battery contains one cell, while a 6V or 9V battery will contain multiple cells connected in series. The total voltage of the battery is the sum of the voltages of its cells.

Electrode Terminology

The electrodes of a cell are often referred to as the *anode* and the *cathode*. These terms are confusing because the electrons enter the anode inside the cell and leave it outside the cell, while electrons enter the cathode from outside the cell

and leave it inside the cell. Thus, the anode is an electron emitter if you look at it externally, but the cathode is an electron emitter if you look at it internally.

Conventional current is imagined to flow in the opposite direction to electrons, and therefore, outside the cell, this current flows from the cathode to the anode, and from this perspective, the cathode can be thought of as being “more positive” than the anode. To remember this, think of the letter t in “cathode” as being a + sign, thus: ca+hode. In larger batteries, the cathode is often painted or tagged red, while the anode may be painted or tagged black or blue.

When a reusable battery is recharged, the flow of electrons reverses and the anode and the cathode effectively trade places. Recognizing this, the manufacturers of rechargeable batteries may refer to the more-positive terminal as the anode. This creates additional confusion, exacerbated further still by electronics manufacturers using the term “cathode” to identify the end of a **diode** which must be “more negative” (i.e., at a lower potential) than the opposite end.

To minimize the risk of errors, it is easiest to avoid the terms “anode” and “cathode” when referring to batteries, and speak instead of the negative and positive terminals. This encyclopedia uses the common convention of reserving the term “cathode” to identify the “more negative” end of any type of diode.

Variants

Three types of batteries exist.

1. *Disposable batteries*, properly (but infrequently) referred to as *primary cells*. They are not reliably rechargeable because their chemical reactions are not easily reversible.
2. *Rechargeable batteries*, properly (but infrequently) known as *secondary cells*. They can be recharged by applying a voltage between

the terminals from an external source such as a [battery charger](#). The materials used in the battery, and the care with which the battery is maintained, will affect the rate at which chemical degradation of the electrodes gradually occurs as it is recharged repeatedly. Either way, the number of charge/discharge cycles is limited.

3. [Fuel Cells](#) require an inflow of a reactive gas such as hydrogen to maintain an electrochemical reaction over a long period. They are beyond the scope of this encyclopedia.

A large **capacitor** may be substituted for a battery for some applications, although it has a lower energy density and will be more expensive to manufacture than a battery of equivalent power storage. A capacitor charges and discharges much more rapidly than a battery because no chemical reactions are involved, but a battery sustains its voltage much more successfully during the discharge cycle. See [Figure 2-5](#).

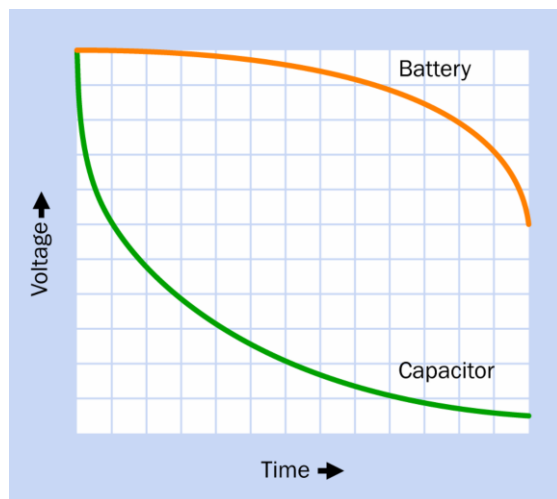


Figure 2-5. The voltage drop of a discharging capacitor is much steeper initially than that of a battery, making capacitors unsuitable as a battery substitute in many applications. However, the ability of a capacitor to discharge very rapidly at high amperage can sometimes be a significant advantage.

Capacitors that can store a very large amount of energy are often referred to as [supercapacitors](#).

Disposable Batteries

The energy density of any disposable battery is higher than that of any type of rechargeable battery, and it will have a much longer shelf life because it loses its charge more slowly during storage (this is known as the [self-discharge rate](#)). Disposable batteries may have a useful life of five years or more, making them ideal for applications such as smoke detectors, handheld remotes for consumer electronics, or emergency flashlights.

Disposable batteries are not well suited to delivering high currents through loads below 75Ω. Rechargeable batteries are preferable for higher-current applications. The bar chart in [Figure 2-6](#) shows the rated and actual capabilities of an alkaline battery relative to the three most commonly used rechargeable types, when the battery is connected with a resistance that is low enough to assure complete discharge in 1 hour.

The manufacturer's rating of watt hours per kilo is typically established by testing a battery with a relatively high-resistance load and slow rate of discharge. This rating will not apply in practice if a battery is discharged with a C-rate of 1, meaning complete discharge during 1 hour.

Common types of disposable batteries are [zinc-carbon cells](#) and [alkaline cells](#). In a zinc-carbon cell, the negative electrode is made of zinc while the positive electrode is made of carbon. The limited power capacity of this type of battery has reduced its popularity, but because it is the cheapest to manufacture, it may still be found where a company sells a product with "batteries included." The electrolyte is usually ammonium chloride or zinc chloride. The 9V battery in [Figure 2-7](#) is actually a zinc-carbon battery according to its supplier, while the smaller one beside it is a 12V alkaline battery designed for use in burglar alarms. These examples show that batteries cannot always be identified correctly by a casual assessment of their appearance.

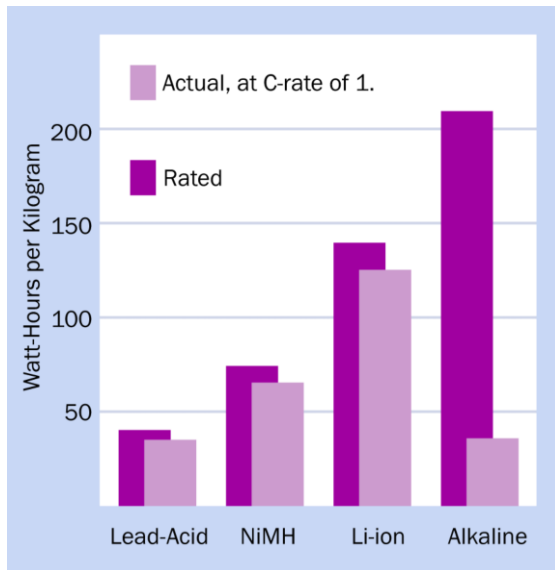


Figure 2-6. Because of their relatively high internal resistance, alkaline batteries are especially unsuited to high discharge rates, and should be reserved for applications where a small current is required over a long period. (Chart derived from <http://batteryuniversity.com>.)



Figure 2-7. At left, a cheap carbon-zinc battery; at right, a 12V alkaline burglar-alarm battery. See text for additional details.

In an alkaline cell, the negative electrode is made of zinc powder, the positive electrode is manganese dioxide, and the electrolyte is potassium hydroxide. An alkaline cell may provide between three to five times the power capacity of an equal size of zinc-carbon cell and is less susceptible to voltage drop during the discharge cycle.

Extremely long shelf life is necessary in some military applications. This may be achieved by using a *reserve battery*, in which the internal chemical compounds are separated from each other but can be recombined prior to use.

Rechargeable Batteries

Commonly used types are *lead-acid*, *nickel cadmium* (abbreviated *NiCad* or *NiCd*), *nickel-metal hydride* (abbreviated *NiMH*), *lithium-ion* (abbreviated *Li-ion*), and *lithium-ion polymer*.

Lead-acid batteries have existed for more than a century and are still widely used in vehicles, burglar alarms, emergency lighting, and large power backup systems. The early design was described as *flooded*; it used a solution of sulfuric acid (generically referred to as *battery acid*) as its electrolyte, required the addition of distilled water periodically, and was vented to allow gas to escape. The venting also allowed acid to spill if the battery was tipped over.

The *valve-regulated lead-acid* battery (*VRLA*) has become widely used, requiring no addition of water to the cells. A pressure relief valve is included, but will not leak electrolyte, regardless of the position of the battery. VRLA batteries are preferred for *uninterruptible power supplies* for data-processing equipment, and are found in automobiles and in electric wheelchairs, as their low gas output and security from spillage increases their safety factor.

VRLA batteries can be divided into two types: absorbed glass mat (AGM) and gel batteries. The electrolyte in an AGM is absorbed in a fiber-glass mat separator. In a gel cell, the electrolyte is mixed with silica dust to form an immobilized gel.

The term *deep cycle battery* may be applied to a lead-acid battery and indicates that it should be more tolerant of discharge to a low level—perhaps 20 percent of its full charge (although manufacturers may claim a lower number). The plates in a standard lead-acid battery are composed of a lead *sponge*, which maximizes the surface area available to acid in the battery but can be phys-

ically abraded by deep discharge. In a deep cycle battery, the plates are solid. This means they are more robust, but are less able to supply high amperage. If a deep-discharge battery is used to start an internal combustion engine, the battery should be larger than a regular lead-acid battery used for this purpose.

A sealed lead-acid battery intended to power an external light activated by a motion detector is shown in [Figure 2-8](#). This unit weighs several pounds and is trickle-charged during the day-time by a 6" × 6" solar panel.



Figure 2-8. A lead-acid battery from an external light activated by a motion sensor.

Nickel-cadmium (*NiCad*) batteries can withstand extremely high currents, but have been banned in Europe because of the toxicity of metallic cadmium. They are being replaced in the United States by *nickel-metal hydride* (*NiMH*) types, which are free from the *memory effect* that can prevent a NiCad cell from fully recharging if it has been left for weeks or months in a partially discharged state.

Lithium-ion and lithium-ion polymer batteries have a better energy-to-mass ratio than NiMH batteries, and are widely used with electronic devices such as laptop computers, media players, digital cameras, and cellular phones. Large arrays of lithium batteries have also been used in some electric vehicles.

Various small rechargeable batteries are shown in [Figure 2-9](#). The NiCad pack at top-left was manufactured for a cordless phone and is rapidly becoming obsolete. The 3V lithium battery at top-right was intended for a digital camera. The three batteries in the lower half of the photograph are all rechargeable NiMH substitutes for 9V, AA, and AAA batteries. The NiMH chemistry results in the AA and AAA single-cell batteries being rated for 1.2V rather than 1.5V, but the manufacturer claims they can be substituted for 1.5V alkaline cells because NiMH units sustain their rated voltage more consistently over time. Thus, the output from a fresh NiMH battery may be comparable to that of an alkaline battery that is part-way through its discharge cycle.



Figure 2-9. Top left: NiCad battery pack for a cordless phone. Top right: Lithium battery for a digital camera. The other batteries are rechargeable NiMH substitutes for everyday alkaline cells.

NiMH battery packs are available to deliver substantial power while being smaller and lighter than lead-acid equivalents. The NiMH package in [Figure 2-10](#) is rated for 10Ah, and consists of ten

D-size NiMH batteries wired in series to deliver 12VDC. This type of battery pack is useful in robotics and other applications where a small motor-driven device must have free mobility.

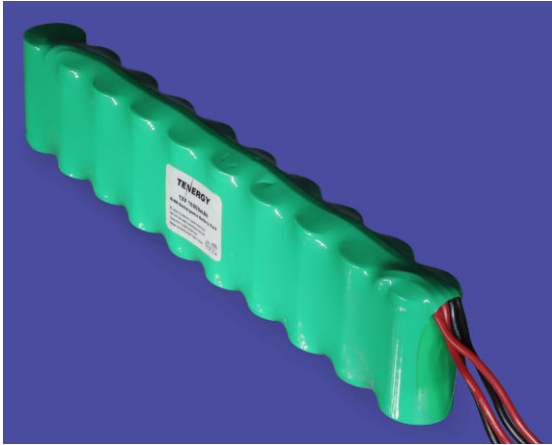


Figure 2-10. This NiMH battery pack is rated at 10Ah and delivers 12 volts from ten D-size cells wired in series.

Values

Amperage

The current delivered by a battery will be largely determined by the resistance of the external load placed between its terminals. However, because ion transfer must occur inside the battery to complete the circuit, the current will also be limited by the *internal resistance* of the battery. This should be thought of as an active part of the circuit.

Since a battery will deliver no current if there is no load, current must be measured while a load is attached, and cannot be measured by a meter alone. The meter will be immediately overloaded, with destructive results, if it is connected directly between the terminals of a battery, or in parallel with the load. Current must always be measured with the meter in series with the load, and the polarity of the meter must correspond with the polarity of the battery. See [Figure 2-11](#).

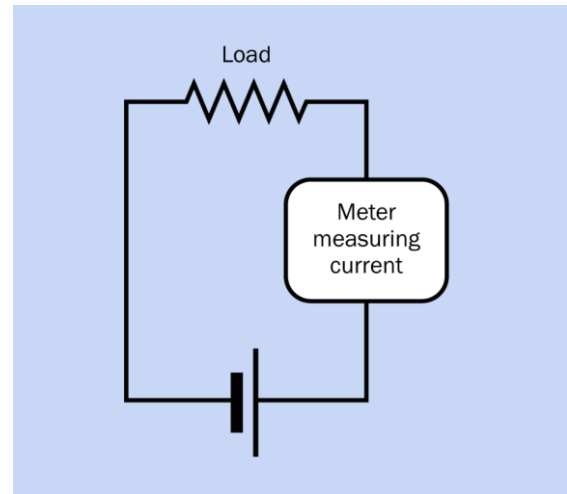


Figure 2-11. When measuring current using an ammeter (or a multimeter configured to measure amps), the meter must be placed in series with the battery and a load. To avoid damaging the meter, it must never be applied directly across the terminals of the battery, or in parallel with a load. Be careful to observe the polarity of the meter.

Capacity

The *electrical capacity* of a battery is measured in *amp-hours*, abbreviated *Ah*, *AH*, or (rarely) *A/H*. Smaller values are measured in *milliamp-hours*, usually abbreviated *mAh*. If *I* is the current being drawn from a battery (in amps) and *T* is the time for which the battery can deliver that current (in hours), the amp-hour capacity is given by the formula:

$$Ah = I * T$$

By turning the formula around, if we know the amp-hour rating that a manufacturer has determined for a battery, we can calculate the time in hours for which a battery can deliver a particular current:

$$T = Ah / I$$

Theoretically, *Ah* is a constant value for any given battery. Thus a battery rated for 4Ah should provide 1 amp for 4 hours, 4 amps for 1 hour, 5 amps for 0.8 hours (48 minutes), and so on.

In reality, this conveniently linear relationship does not exist. It quickly breaks down as the

current rises, especially when using lead-acid batteries, which do not perform well when required to deliver high current. Some of the current is lost as heat, and the battery may be electrochemically incapable of keeping up with demand.

The *Peukert number* (named after its German originator in 1897) is a fudge factor to obtain a more realistic value for T at higher currents. If n is the Peukert number for a particular battery, then the previous formula can be modified thus:

$$T = Ah / I^n$$

Manufacturers usually (but not always) supply Peukert's number in their specification for a battery. So, if a battery has been rated at 4Ah, and its Peukert number is 1.2 (which is typical for lead-acid batteries), and I=5 (in other words, we want to know for how long a time, T, the battery can deliver 5 amps):

$$T = 4 / 5^{1.2} = \text{approximately } 4 / 6.9$$

This is about 0.58 hours, or 35 minutes—much less than the 48 minutes that the original formula suggested.

Unfortunately, there is a major problem with this calculation. In Peukert's era, the amp-hour rating for a battery was established by a manufacturer by drawing 1A and measuring the time during which the battery was capable of delivering that current. If it took 4 hours, the battery was rated at 4Ah.

Today, this measurement process is reversed. Instead of specifying the current to be drawn from the battery, a manufacturer specifies the time for which the test will run, then finds the maximum current the battery can deliver for that time. Often, the time period is 20 hours. Therefore, if a battery has a modern 4Ah rating, testing has probably determined that it delivered 0.2A for 20 hours, not 1A for 4 hours, which would have been the case in Peukert's era.

This is a significant distinction, because the same battery that can deliver 0.2A for 20 hours will not

be able to satisfy the greater demand of 1A for 4 hours. Therefore the old amp-hour rating and the modern amp-hour rating mean different things and are incompatible. If the modern Ah rating is inserted into the old Peukert formula (as it was above), the answer will be misleadingly optimistic. Unfortunately, this fact is widely disregarded. Peukert's formula is still being used, and the performance of many batteries is being evaluated incorrectly.

The formula has been revised (initially by Chris Gibson of SmartGauge Electronics) to take into account the way in which Ah ratings are established today. Suppose that AhM is the modern rating for the battery's capacity in amp-hours, H is the duration in hours for which the battery was tested when the manufacturer calibrated it, n is Peukert's number (supplied by the manufacturer) as before, and I is the current you hope to draw from the battery. This is the revised formula to determine T:

$$T = H * (AhM / (I * H)^n)$$

How do we know the value for H? Most (not all) manufacturers will supply this number in their battery specification. Alternatively, and confusingly, they may use the term *C-rate*, which can be defined as 1/H. This means you can easily get the value for H if you know the C-rate:

$$H = 1 / \text{C-rate}$$

We can now use the revised formula to rework the original calculation. Going back to the example, if the battery was rated for 4Ah using the modern system, in a discharge test that lasted 20 hours (which is the same as a C-rate of 0.05), and the manufacturer still states that it has a Peukert number of 1.2, and we want to know for how long we can draw 5A from it:

$$T = 20 * (4 / (5 * 20)^{1.2}) = \text{approximately } 20 * 0.021$$

This is about 0.42 hours, or 25 minutes—quite different from the 35 minutes obtained with the old version of the formula, which should never be used when calculating the probable dis-

charge time based on a modern Ah rating. These issues may seem arcane, but they are of great importance when assessing the likely performance of battery-powered equipment such as electric vehicles.

Figure 2-12 shows the probable actual performance of batteries with Peukert numbers of 1.1, 1.2, and 1.3. The curves were derived from the revised version of Peukert's formula and show how the number of amp-hours that you can expect diminishes for each battery as the current increases. For example, if a battery that the manufacturer has assigned a Peukert number of 1.2 is rated at 100Ah using the modern 20-hour test, but we draw 30A from it, the battery can actually deliver only 70Ah.

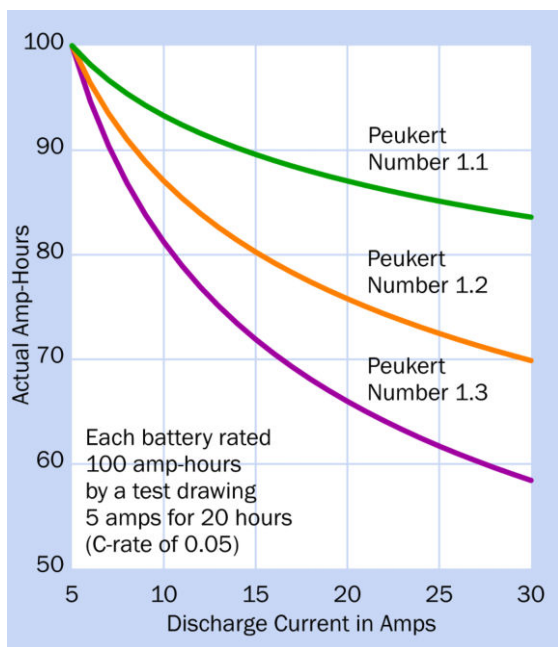


Figure 2-12. Actual amp-hour performance that should be expected from three batteries of Peukert numbers 1.1, 1.2, and 1.3 when they discharge currents ranging from 5 to 30 amps, assuming that the manufacturer has rated each battery at 100Ah using the modern system, which usually entails a 20-hour test (a C-rate of 0.05).

One additional factor: For any rechargeable battery, the Peukert number gradually increases with age, as the battery deteriorates chemically.

Voltage

The rated voltage of a fully charged battery is known as the *open circuit voltage* (abbreviated *OCV* or V_{oc}), defined as the potential that exists when no load is imposed between the terminals. Because the internal resistance of a volt meter (or a multimeter, when it is used to measure DC volts) is very high, it can be connected directly between the battery terminals with no other load present, and will show the OCV quite accurately, without risk of damage to the meter. A fully charged 12-volt car battery may have an OCV of about 12.6 volts, while a fresh 9-volt alkaline battery typically has an OCV of about 9.5 volts. Be extremely careful to set a multimeter to measure DC volts before connecting it across the battery. Usually this entails plugging the wire from the red probe into a socket separately reserved for measuring voltage, not amperage.

The voltage delivered by a battery will be pulled down significantly when a load is applied to it, and will decrease further as time passes during a discharge cycle. For these reasons, a **voltage regulator** is required when a battery powers components such as digital integrated circuit chips, which do not tolerate a wide variation in voltage.

To measure voltage while a load is applied to the battery, the meter must be connected in parallel with the load. See Figure 2-13. This type of measurement will give a reasonably accurate reading for the potential applied to the load, so long as the resistance of the load is relatively low compared with the internal resistance of the meter.

Figure 2-14 shows the performance of five commonly used sizes of alkaline batteries. The ratings in this chart were derived for alkaline batteries under favorable conditions, passing a small current through a relatively high-ohm load for long periods (40 to 400 hours, depending on battery type). The test continued until the final voltage for each 1.5V battery was 0.8V, and the final voltage for the 9V battery was a mere 4.8V. These voltages were considered acceptable when the

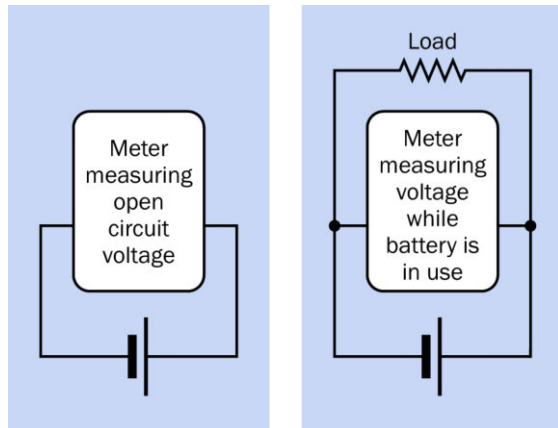


Figure 2-13. When using a volt meter (or a multimeter configured to measure voltage), the meter can be applied directly between the battery terminals to determine the open-circuit voltage (OCV), or in parallel with a load to determine the voltage actually supplied during use. A multimeter must be set to measure DC volts before connecting it across a battery. Any other setting may damage the meter.

Ah ratings for the batteries were calculated by the manufacturer, but in real-world situations, a final voltage of 4.8V from a 9V battery is likely to be unacceptable in many electronics applications.

Battery type	Rating (Ah)	Final voltage	Load (ohms)	Current (mA)
AAA	1.15	0.8	75	20
AA	2.87	0.8	75	20
C	7.8	0.8	39	40
D	17	0.8	39	40
9V	0.57	4.8	620	14

Figure 2-14. The voltage delivered by a battery may drop to a low level while a manufacturer is establishing an amp-hour rating. Values for current, shown in the chart, were calculated subsequently as estimated averages, and should be considered approximate. (Derived from a chart published by Panasonic.)

As a general rule of thumb, if an application does not tolerate a significant voltage drop, the manufacturer's amp-hour rating for a small battery may be divided by 2 to obtain a realistic number.

How to Use it

When choosing a battery to power a circuit, considerations will include the intended shelf life, maximum and typical current drain, and battery weight. The amp-hour rating of a battery can be used as a very approximate guide to determine its suitability. For 5V circuits that impose a drain of 100mA or less, it is common to use a 9V battery, or six 1.5V batteries in series, passing current through a **voltage regulator** such as the LM7805. Note that the voltage regulator requires energy to function, and thus it imposes a voltage drop that will be dissipated as heat. The minimum drop will vary depending on the type of regulator used.

Batteries or cells may be used in series or in parallel. In series, the total voltage of the chain of cells is found by summing their individual voltages, while their amp-hour rating remains the same as for a single cell, assuming that all the cells are identical. Wired in parallel, the total voltage of the cells remains the same as for a single cell, while the combined amp-hour value is found by summing their individual amp-hour ratings, assuming that all the batteries are identical. See [Figure 2-15](#).

In addition to their obvious advantage of portability, batteries have an additional advantage of being generally free from power spikes and noise that can cause sensitive components to misbehave. Consequently, the need for smoothing will depend only on possible noise created by other components in the circuit.

Motors or other inductive loads draw an initial surge that can be many times the current that they use after they start running. A battery must be chosen that will tolerate this surge without damage.

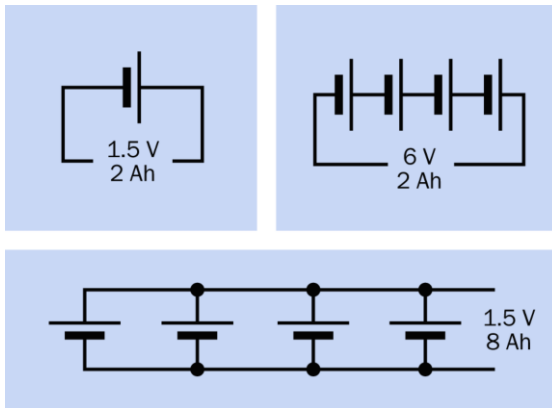


Figure 2-15. Theoretical results of using 1.5V cells in series or in parallel, assuming a 2Ah rating for one cell.

Because of the risk of fire, United States airline regulations limit the amp-hour capacity of lithium-ion batteries in any electronic device in carry-on or checked passenger baggage. If a device may be carried frequently as passenger baggage (for example, emergency medical equipment), NiMH batteries are preferred.

What Can Go Wrong

Short Circuits: Overheating and Fire

A battery capable of delivering significant current can overheat, catch fire, or even explode if it is short-circuited. Dropping a wrench across the terminals of a car battery will result in a bright flash, a loud noise, and some molten metal. Even a 1.5-volt alkaline AA battery can become too hot to touch if its terminals are shorted together. (Never try this with a rechargeable battery, which has a much lower internal resistance, allowing much higher flow of current.) Lithium-ion batteries are particularly dangerous, and almost always are packaged with a current-limiting component that should not be disabled. A short-circuited lithium battery can explode.

If a battery pack is used as a cheap and simple workbench DC power supply, a **fuse** or circuit breaker should be included. Any device that uses significant battery power should be fused.

Diminished Performance Caused by Improper Recharging

Many types of batteries require a precisely measured charging voltage and a cycle that ends automatically when the battery is fully charged. Failure to observe this protocol can result in chemical damage that may not be reversible. A charger should be used that is specifically intended for the type of battery. A detailed comparison of chargers and batteries is outside the scope of this encyclopedia.

Complete Discharge of Lead-Acid Battery

Complete or near-complete discharge of a lead-acid battery will significantly shorten its life (unless it is specifically designed for deep-cycle use—although even then, more than an 80% discharge is not generally recommended).

Inadequate Current

Chemical reactions inside a battery occur more slowly at low temperatures. Consequently, a cold battery cannot deliver as much current as a warm battery. For this reason, in winter weather, a car battery is less able to deliver high current. At the same time, because engine oil becomes more viscous as the temperature falls, the starter motor will demand more current to turn the engine. This combination of factors explains the tendency of car batteries to fail on cold winter mornings.

Incorrect Polarity

If a battery charger or generator is connected with a battery with incorrect polarity, the battery may experience permanent damage. The **fuse** or **circuit breaker** in a charger may prevent this from occurring and may also prevent damage to the charger, but this cannot be guaranteed.

If two high-capacity batteries are connected with opposite polarity (as may happen when a clumsy attempt is made to start a stalled car with jumper cables), the results may be explosive. Never lean over a car battery when attaching cables to it, and ideally, wear eye protection.

Reverse Charging

Reverse charging can occur when a battery becomes completely discharged while it is wired (correctly) in series with other batteries that are still delivering current. In the upper section of the schematic at [Figure 2-16](#) two healthy 6V batteries, in series, are powering a resistive load. The battery on the left applies a potential of 6 volts to the battery on the right, which adds its own 6 volts to create a full 12 volts across the load. The red and blue lines indicate volt meter leads, and the numbers show the reading that should be observed on the meter.

In the second schematic, the battery on the left has become exhausted and is now a “dead weight” in the circuit, indicated by its gray color. The battery on the right still sustains a 6-volt potential. If the internal resistance of the dead battery is approximately 1 ohm and the resistance of the load is approximately 20 ohms, the potential across the dead battery will be about 0.3 volts, in the opposite direction to its normal charged voltage. Reverse charging will result and can damage the battery. To avoid this problem, a battery pack containing multiple cells should never be fully discharged.

Sulfurization

When a lead-acid battery is partially or completely discharged and is allowed to remain in that state, sulfur tends to build up on its metal plates. The sulfur gradually tends to harden, forming a barrier against the electrochemical reactions that are necessary to recharge the battery. For this reason, lead-acid batteries should not be allowed to sit for long periods in a discharged condition. Anecdotal evidence suggests that even a very small trickle-charging current can prevent sulfurization, which is why some people recommend attaching a small solar panel to a battery that is seldom used—for example, on a sail boat, where the sole function of the battery is to start an auxiliary engine when there is insufficient wind.

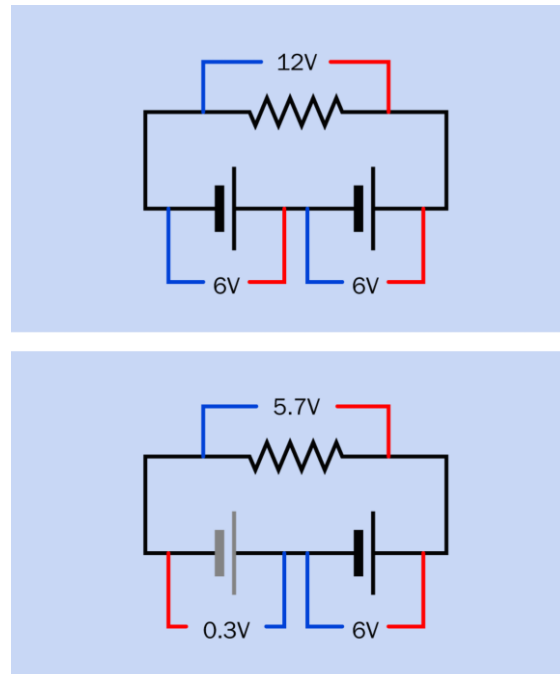


Figure 2-16. When a pair of 6V batteries is placed in series to power a resistive load, if one of the batteries discharges completely, it becomes a load instead of a power source, and will be subjected to reverse charging, which may cause permanent damage.

High Current Flow Between Parallel Batteries

If two batteries are connected in parallel, with correct polarity, but one of them is fully charged while the other is not, the charged battery will attempt to recharge its neighbor. Because the batteries are wired directly together, the current will be limited only by their internal resistance and the resistance of the cables connecting them. This may lead to overheating and possible damage. The risk becomes more significant when linking batteries that have high Ah ratings. Ideally they should be protected from one another by high-current **fuses**.

jumper

3

A jumper may also be referred to as a *jumper socket* or a *shunt*. A jumper should not be confused with *jumper wires*, which are not considered components for the purposes of this encyclopedia.

OTHER RELATED COMPONENTS

- **switch** (See [Chapter 6](#))

What It Does

A jumper is a low-cost substitute for a **switch**, where a connection has to be made (or unmade) only a few times during the lifetime of a product. Typically it allows a function or feature on a circuit board to be set on a semipermanent basis, often at the time of manufacture. A *DIP switch* performs the same function more conveniently. See "[DIP](#)" ([page 43](#)).

There is no standardized schematic symbol to represent a jumper.

How It Works

A jumper is a very small rectangular plastic tab containing two (or sometimes more) metal sockets usually spaced either 0.1" or 2mm apart. The sockets are connected electrically inside the tab, so that when they are pushed over two (or more) pins that have been installed on a circuit board for this purpose, the jumper shorts the pins together. The pins are usually 0.025" square and are often part of a *header* that is soldered into the board. In a parts catalogue, jumpers may be found in a section titled "Headers and Wire Housings" or similar.

Three jumpers are shown in [Figure 3-1](#). The blue one contains two sockets spaced 0.1" and is deep

enough to enclose the pins completely. The red one contains two sockets spaced 2mm and may allow the tips of the pins to emerge from its opposite end. The black one contains four sockets, each pair spaced 0.1" apart.

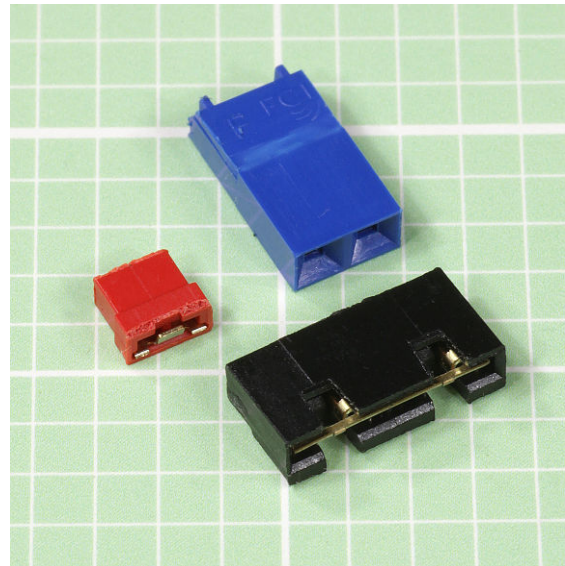


Figure 3-1. Three jumpers containing two sockets spaced 2mm (left), two sockets spaced 0.1" (top right), and four sockets, each pair spaced 0.1" (bottom right).

The set of pins with which a jumper is used is often referred to as a *header*. Headers are available with pins in single or dual rows. Some headers are designed to be snapped off to provide the desired number of pins. A dual 28-pin header is shown in [Figure 3-2](#) with a black jumper pushed onto a pair of pins near the midpoint.

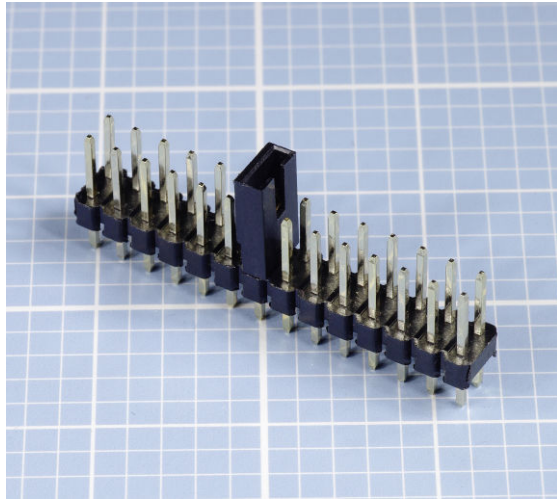


Figure 3-2. A jumper pushed onto a pair of pins midway along a dual 28-pin header.

Variants

A *jumper assembly* may be a kit containing not only the jumper but also the array of pins with which it is intended to be used. Check the manufacturer's datasheet to find out exactly what is included.

The most common types of jumpers have two sockets only, but variants are available with as many as 12 sockets, which may be arranged in one or two rows. *Header sockets* may be used as a substitute for purpose-made jumpers, with the advantage that they are often sold in long strips that can be snapped off to provide as many sockets as needed. However, the pins attached to header sockets must be manually connected by soldering small lengths of wire between them.

In some jumpers, the plastic tab extends upward for about half an inch and functions as a finger grip, making the jumper much easier to hold during insertion and removal. This is a desirable feature if there is room to accommodate it.

The sockets inside a jumper are often made from phosphor-bronze, copper-nickel alloy, tin alloy, or brass alloy. They are usually gold-plated, but in some instances are tin-plated.

Rarely, a jumper may consist of a metal strip with U-shaped connections suitable for being used in conjunction with screw terminals. Two jumpers of this type are shown in [Figure 3-3](#). They should not be confused with high-amperage fuses that look superficially similar.

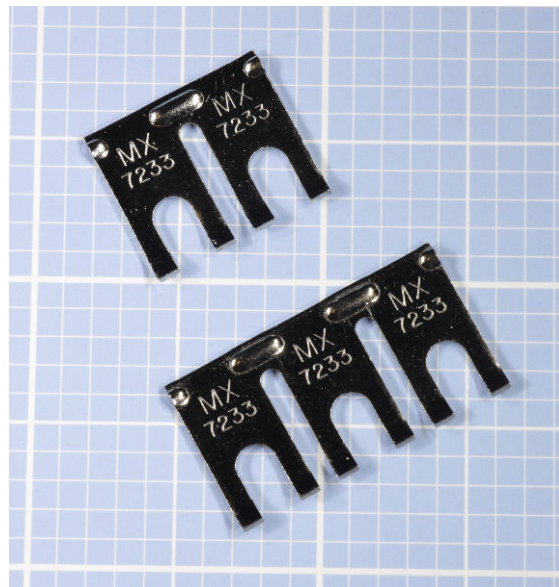


Figure 3-3. These jumpers are designed to short together two or three screw terminals.

Values

The spacing between the sockets in a jumper is referred to as its *pitch*. As previously noted, 0.1" and 2mm are the most popular values.

A typical maximum rating for a jumper of 0.1" pitch is 2A or 2.5A at 250V.

How to Use it

A jumper may activate a “set it and forget it” circuit function. An example would be the factory configuration of a product to work with 115VAC or 230VAC power input. End users were expected to set jumpers in some computer equipment sold during the 1980s, but this is no longer the case.

What Can Go Wrong

Jumpers are easily dropped, easily lost, and easily placed incorrectly. When purchasing jumpers, buy extras to compensate for their fragility and the ease of losing them.

Any location where a jumper may be used should be clearly labelled to define the function of each setting.

Cheap, poorly made jumpers may self-destruct from mechanical stresses when removed from their pins. The plastic casing can come away, leaving the sockets clinging naked to the pins protruding from the circuit board. This is another reason why it is a good idea to have a small stock of spare jumpers for emergencies.

Oxidation in jumpers where the contacts are not gold- or silver-plated can create electrical resistance or unreliable connections.

fuse

4

The alternate spelling *fuze* is seldom used.

OTHER RELATED COMPONENTS

- None

What It Does

A fuse protects an electrical circuit or device from excessive current when a metal element inside it melts to create an open circuit. With the exception of *resettable fuses* (discussed separately in “Resettable Fuses” (page 24)), a fuse must be discarded and replaced after it has fulfilled its function.

When high current melts a fuse, it is said to *blow* or *trip* the fuse. (In the case of a resettable fuse, only the word *trip* is used.)

A fuse can work with either AC or DC voltage, and can be designed for almost any current. In residential and commercial buildings, *circuit breakers* have become common, but a large cartridge fuse may still be used to protect the whole system from short-circuits or from overcurrent caused by lightning strikes on exposed power lines.

In electronic devices, the **power supply** is almost always fused.

Schematic symbols for a fuse are shown in Figure 4-1. Those at the right and second from right are most frequently used. The one in the center is approved by ANSI, IEC, and IEEE but is seldom seen. To the left of that is the fuse symbol understood by electrical contractors in architectural plans. The symbol at far left used to be common but has fallen into disuse.

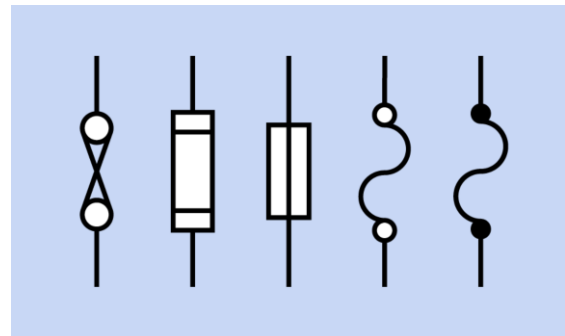


Figure 4-1. Alternate schematic symbols for a fuse. See text for explanation.

How It Works

The *element* in a fuse is usually a wire or thin metal strip mounted between two terminals. In a cartridge fuse, it is enclosed in a glass or ceramic cylinder with a contact at each end, or in a small metallic can. (Old-style, large, high-amperage fuses may be packaged in a paper or cardboard tube.) The traditional glass cartridge allows visual inspection to confirm that the fuse has blown.

A fuse responds only to current, not to voltage. When choosing a fuse that will be reliable in conditions of steady current consumption, a safe rule is to figure the maximum amperage when all components are functioning and add 50%. However, if current surges or spikes are likely, their duration will be relevant. If *I* is the current surge

in amps and t is its duration in seconds, the surge sensitivity of a fuse—which is often referred to verbally or in printed format as I^2t —is given by the formula:

$$I^2t = I^2 * t$$

Some semiconductors also have an I^2t rating, and should be protected with a similarly rated fuse.

Any fuse will present some resistance to the current flowing through it. Otherwise, the current would not generate the heat that blows the fuse. Manufacturer datasheets list the voltage drop that the internal resistance of a fuse is likely to introduce into a circuit.

Values

The *current rating* or *rated current* of a fuse is usually printed or stamped on its casing, and is the maximum flow that it should withstand on a continuous basis, at the ambient temperature specified by the manufacturer (usually 25 degrees Centigrade). The ambient temperature refers to the immediate environment of the fuse, not the larger area in which it may be located. Note that in an enclosure containing other components, the temperature is usually significantly higher than outside the enclosure.

Ideally a fuse should function reliably and indefinitely at its rated maximum amperage, but should blow just as reliably if the current rises by approximately 20% beyond the maximum. In reality, manufacturers recommend that continuous loading of a fuse should not exceed 75% of its rating at 25 degrees Centigrade.

The *voltage rating* or *rated voltage* of a fuse is the maximum voltage at which its element can be counted on to melt in a safe and predictable manner when it is overloaded by excess current. This is sometimes known as the *breaking capacity*. Above that rating, the remaining pieces of the fuse element may form an arc that sustains some electrical conduction.

A fuse can always be used at a lower voltage than its rating. If it has a breaking capacity of 250V, it will still provide the same protection if it is used at 5V.

Four differently rated glass cartridge fuses are shown in [Figure 4-2](#). The one at the top is a slow-blowing type, rated at 15A. Its element is designed to absorb heat before melting. Below it is a 0.5A fuse with a correspondingly thinner element. The two smaller fuses are rated at 5A each. The center two fuses have a maximum voltage rating of 250V, while the one at the top is rated at 32V and the one at the bottom is rated at 350V. Clearly, the size of a fuse should never be used as a guide to its ratings.

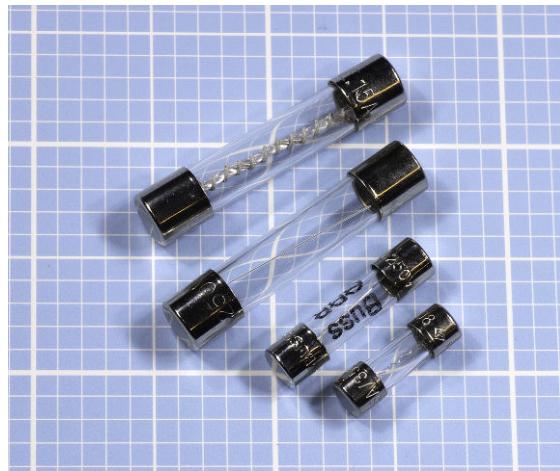


Figure 4-2. Four glass cartridge fuses. See text for details.

Variants

Early power fuses in residential buildings consisted of bare nichrome wire wrapped around a porcelain holder. In the 1890s, Edison developed *plug fuses* in which the fuse was contained in a porcelain module with a screw thread, compatible with the base of an incandescent bulb. This design persisted in some U.S. urban areas for more than 70 years, is still found in old buildings, and is still being manufactured.

Small Cartridge Fuses

Small cartridge fuses for appliances and electronics equipment—such as those shown in Figure 4-2—are available in sizes tabulated in Figure 4-3. With the exception of the 4.5mm diameter fuse (a European addition), these sizes were originally measured in inches; today, they are often described only with the equivalent metric measurement. Any cartridge fuse is usually available with the option of a lead attached to it at each end, so that it can be used as a through-hole component.

Fuse type	Diameter (inches)	Diameter (metric)	Length (inches)	Length (metric)
1AG	1/4"	6mm	5/8"	16mm
2AG	0.177"	4.5mm	0.588"	15mm
3AG	1/4"	6mm	1-1/4"	32mm
4AG	9/32"	7mm	1-1/4"	32mm
5AG	13/32"	10mm	1-1/2"	38mm
7AG	1/4"	6mm	7/8"	22mm
8AG	1/4"	6mm	1"	25mm

Figure 4-3. The approximate physical sizes of commonly used small glass or ceramic cartridge fuses are shown here with the codes that are often used to identify them.

Fuses may be fast acting, medium acting, or slow-blowing, the last of which may alternatively be referred to as *delay fuses*. Extra-fast-acting fuses are available from some manufacturers. The term *Slo-Blo* is often used but is actually a trademark of Littelfuse. None of the terms describing the speed of action of a fuse has been standardized with a specific time or time range.

Some cartridge fuses are available in a ceramic format as an alternative to the more common glass cylinder. If accidental application of extremely high current is possible (for example, in a multimeter that can be set to measure amps, and may be accidentally connected across a powerful battery), a ceramic cartridge is prefera-

ble because it contains a filler that will help to stop an arc from forming. Also, if a fuse is physically destroyed by application of very high current, ceramic fragments may be preferable to glass fragments.

Automotive Fuses

Automotive fuses are identifiable by their use of blades designed for insertion in flat sockets where the fuse is unlikely to loosen as a result of vibration or temperature changes. The fuses come in various sizes, and are uniformly color-coded for easy identification.

A selection of automotive fuses is shown in Figure 4-4. The type at the top is typically described as a “maxi-fuse” while the type at bottom-left is a “mini-fuse.” Here again, size is irrelevant to function, as all three of those pictured are rated 30A at 32V.

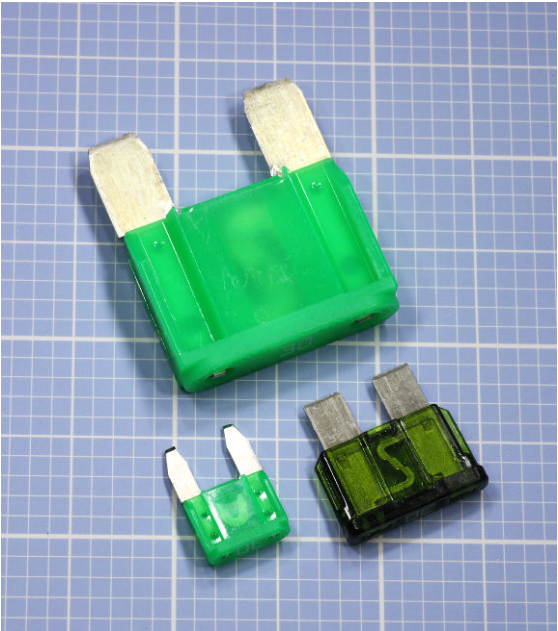


Figure 4-4. Three automotive fuses. All have the same rating: 30A at 32V.

In Figure 4-5, the largest of the fuses from Figure 4-4 has been cut open to reveal its element.

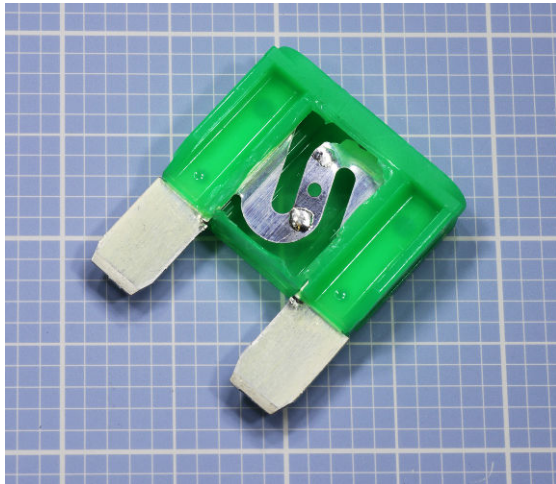


Figure 4-5. The largest fuse from the previous figure, cut open to reveal its element.

Usually automotive fuses are mounted together in a block, but if aftermarket accessory equipment is added, it may be protected by an inline fuse in a holder that terminates in two wires. This is shown with two sample fuses in [Figure 4-6](#). Similar inline fuse holders are manufactured for other types of fuses.

Strip Fuses

High-amperage fuses for vehicles may be sold in “strip fuse” format, also known as a *fusible link*, designed to be clamped between two screw-down terminals. Since some **jumpers** may look very similar, it is important to keep them separate. A strip fuse is shown in [Figure 4-7](#).

Through-Hole Fuses

Small fuses with radial leads, which seem appropriate for through-hole insertion in printed circuit boards, are actually often used in conjunction with appropriate sockets, so that they can be easily replaced. They are described in catalogues as “subminiature fuses” and are typically found in laptop computers and their power supplies, also televisions, battery chargers, and air conditioners. Three examples are shown in [Figure 4-8](#). All have slow-blowing characteristics.



Figure 4-6. Two blade-type fuses, commonly used for automotive applications, shown with an inline fuse holder. The plastic cap, at right, is closed over the holder when a fuse has been installed.

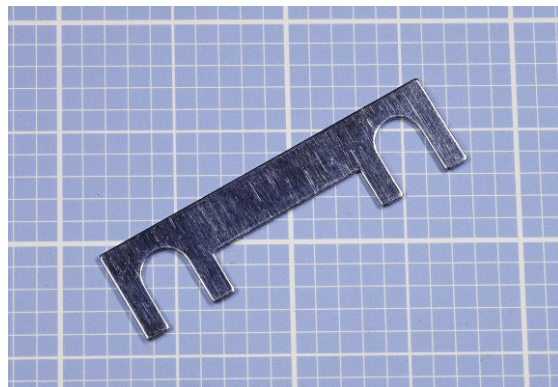


Figure 4-7. This strip fuse is intended for use in diesel vehicles. The example shown is rated 100A at 36V.

Resettable Fuses

Properly known as a *polymeric positive temperature coefficient* fuse (often abbreviated *PTC* or *PPTC*), a *resettable fuse* is a solid-state, encapsulated component that greatly increases its resistance in response to a current overload, but gradually returns to its original condition when the flow of current is discontinued. It can be thought

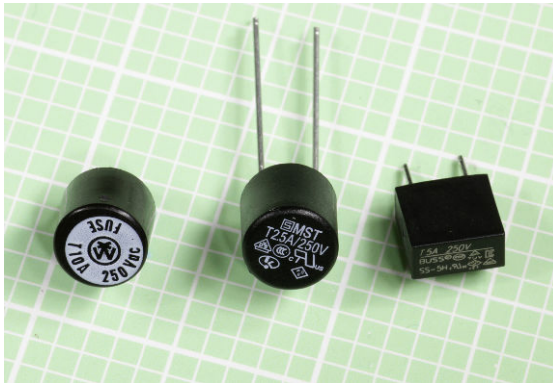


Figure 4-8. Three subminiature fuses terminating in wire leads. From left to right: 10A at 250V, 2.5A at 250V, and 5A at 250V.

of as a **thermistor** that has a nonlinear response. Three through-hole examples are shown in [Figure 4-9](#). While different sizes of cartridge fuse may share the same ratings, differently rated resettable fuses may be identical in size. The one on the left is rated 40A at 30V, while the one on the right is rated 2.5A at 30V. (Note that the codes printed on the fuses are not the same as their manufacturer part numbers.) The fuse at the top is rated 1A at 135V.

When more than the maximum current passes through the fuse, its internal resistance increases suddenly from a few ohms to hundreds of thousands of ohms. This is known as **tripping** the fuse. This inevitably entails a small delay, but is comparable to the time taken for a slow-blowing fuse to respond.

A resettable fuse contains a polymer whose crystalline structure is loaded with graphite particles that conduct electricity. As current flowing through the fuse induces heat, the polymer transitions to an amorphous state, separating the graphite particles and interrupting the conductive pathways. A small current still passes through the component, sufficient to maintain its amorphous state until power is disconnected.

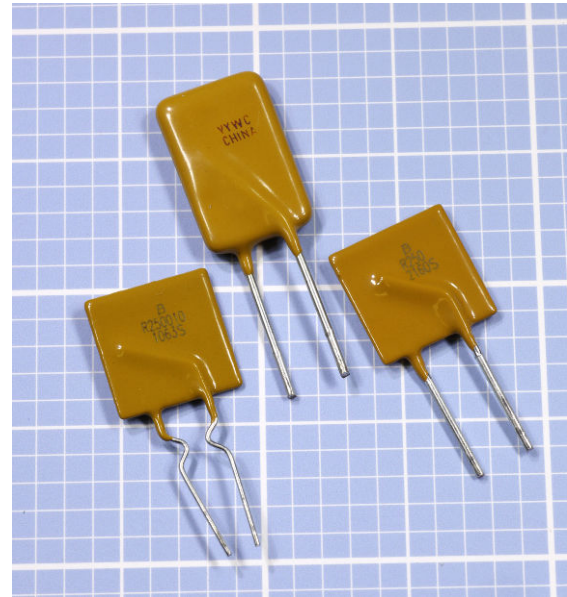


Figure 4-9. Some through-hole resettable fuses. See text for details.

After the resettable fuse cools, it gradually recrystallizes, although its resistance does not fall back completely to its original value for more than an hour.

The maximum safe level of current for a resettable fuse is known as the **hold current**, while the current that triggers its response is termed the **trip current**. Resettable fuses are available with trip-current ratings from 20mA to 100A. While conventional appliance and electronics fuses may be rated as high as 600V, resettable fuses are seldom rated above 100V.

Typical cartridge fuses are affected only to a minor extent by temperature, but the current rating of a resettable fuse may diminish to 75% of its normal value at 50 degrees Centigrade and may drop to 50% of its normal value at 80 degrees Centigrade. In other words, a fuse that is rated for 4A at 25 degrees may tolerate a maximum of only 3A when it operates at twice that temperature. See [Figure 4-10](#).

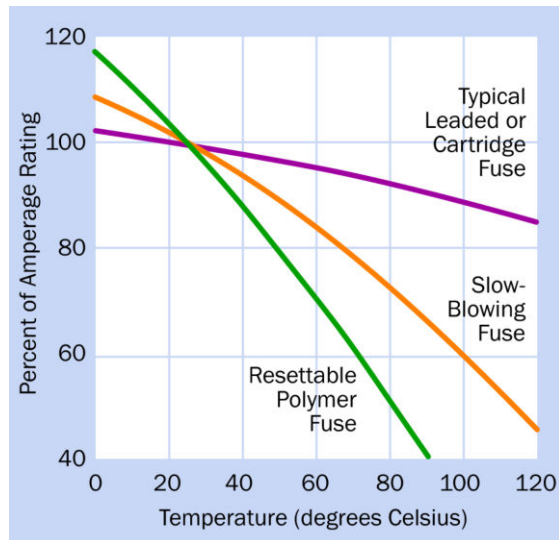


Figure 4-10. The three curves give a very approximate idea of the temperature sensitivity of three types of fuses commonly used to protect electronic equipment. The left-hand scale provides an approximate value for the amperage which will trip the fuse.

Conventional slow-blowing fuses are temperature-sensitive, but to a lesser degree than resettable fuses.

Resettable fuses are used in computer power supplies, USB power sources, and loudspeaker enclosures, where they protect the speaker coils from being overdriven. They are appropriate in situations where a fuse may be tripped relatively often, or where technically unsophisticated users may feel unable to replace a fuse or reset a circuit breaker.

Brand names for resettable fuses include PolySwitch, OptiReset, Everfuse, Polyfuse, and Multifuse. They are available in [surface-mount](#) (SMT) packages or as [through-hole](#) components, but not in cartridge format.

Surface Mount Fuses

Because surface-mount fuses are difficult or impossible to replace after they have been soldered onto the board, they are often resettable.

A surface-mount resettable fuse approximately 0.3" square is shown in [Figure 4-11](#). It is rated for 230V and has an internal resistance of 50 ohms. Its hold current is 0.09A and its trip current is 0.19A.

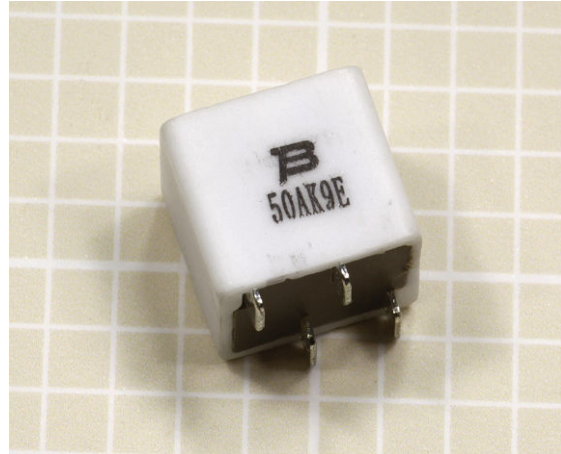


Figure 4-11. A surface-mount resettable fuse. See text for details.

How to Use it

Any equipment that may be plugged into a wall outlet should be fused, not only to protect its components but also to protect users who may open the box and start investigating with a screwdriver.

Equipment that contains powerful motors, pumps, or other inductive loads should be protected with slow-blowing fuses, as the initial surge of current when the equipment is switched on is likely to rise well above the rating of the fuse. A slow-blowing fuse will tolerate a surge for a couple of seconds. Other fuses will not.

Conversely, fast-acting fuses should be used with electronic equipment, especially integrated circuits that are quickly and easily damaged.

Any device using substantial **battery** power should be fused because of the unpredictable and generally bad behavior of batteries when they are short-circuited. Parallel connections between multiple large batteries should be fused

to avoid the possibility that a highly charged battery may attempt to recharge its neighbor(s). Large “J size” fuses rated from 125A to 450A have become common in the solar power community, where banks of lead-acid batteries are often used. These fuses have a thick brass tab at each end, drilled so that they can be bolted into place. Alternatively, they will push-fit into an appropriate [fuseholder](#).

For cartridge fuses up to 1/4” in diameter that don’t have leads attached, appropriately sized fuseholders are available in several formats:

Panel mounted fuse enclosure is probably the most common, consisting of a plastic tube with a spring-contact at the bottom, and a plastic cap with a second contact inside. The cap either screws onto the tube of the fuse, or is pushed down and turned to hold it in place. A nut is provided to secure the fuseholder after it has been inserted into a hole drilled in the panel. The fuse is dropped into the tube, and the cap is applied. This type of holder is available in full-length or shorter, “low profile” formats. A low-profile holder is shown in [Figure 4-12](#). It is shown assembled at right, with its component parts disassembled alongside.



Figure 4-12. A low-profile panel-mounted fuse holder shown disassembled (left) and assembled (right).

Circuit board mounted fuse enclosure is basically the same as the panel-mounted version, but with through-hole solder pins attached.

Fuse block is a small plastic block with two clips on its upper surface for insertion of a cartridge fuse.

Fuse clips can be bought individually, with solder pins for through-hole mounting.

Inline fuse holder is designed to be inserted in a length of wire. Usually made of plastic, it will either terminate it, wires or will have metal contacts to crimp or solder at each end. See [Figure 4-6](#).

Through-hole fuse holders are available for subminiature fuses.

What Can Go Wrong

Repeated Failure

When a fuse in a circuit blows frequently, this is known as [nuisance opening](#). Often it can result from failure to take into account all the aspects of the circuit, such as a large filtering capacitor in a power supply that draws a major surge of current when the power supply is switched on. The formally correct procedure to address this problem is to measure the power surge, properly known as [peak inrush current](#), with an oscilloscope, calculate the $I^2 \cdot t$ of the wave form, and select a fuse with a rating at least 5 times that value.

A fuse should never be replaced with an equivalent length of wire or any other conductor.

Soldering Damage

When a through-hole or surface-mount fuse is soldered into place, heat from the soldering process can cause the soft metal element inside the fuse to melt partially and reflow. This is likely to change the rating of the fuse. Generally, fuses should be treated with the same caution as semiconductors when they are fixed in place with solder.

Placement

A fuse should be placed close to the power source or power input point in a circuit, so that it protects as much of the circuit as possible.

pushbutton

Often referred to as a *pushbutton switch* and sometimes as a *momentary switch*. In this encyclopedia, a pushbutton is considered separately from a **switch**, which generally uses a lever-shaped actuator rather than a button, and has at least one *pole* contact where a pushbutton generally has contacts that are not distinguishable from each other.

OTHER RELATED COMPONENTS

- **switch** (See [Chapter 6](#))
- **rotary switch** (See [Chapter 7](#))

What It Does

A pushbutton contains at least two contacts, which close or open when the button is pressed. Usually a spring restores the button to its original position when external pressure is released. [Figure 5-1](#) shows schematic symbols for pushbuttons. The symbols that share each blue rectangle are functionally identical. At top is a normally-open single-throw pushbutton. At center is a normally-closed single-throw pushbutton. At bottom is a double-throw pushbutton.

Unlike a **switch**, a basic pushbutton does not have a primary contact that can be identified as the *pole*. However, a single pushbutton may close or open two separate pairs of contacts, in which case it can be referred to, a little misleadingly, as a double-pole pushbutton. See [Figure 5-2](#). Different symbols are used for slider pushbuttons with multiple contact pairs; see “[Slider](#)” (page 31).

A generic full-size, two-contact pushbutton is shown in [Figure 5-3](#).

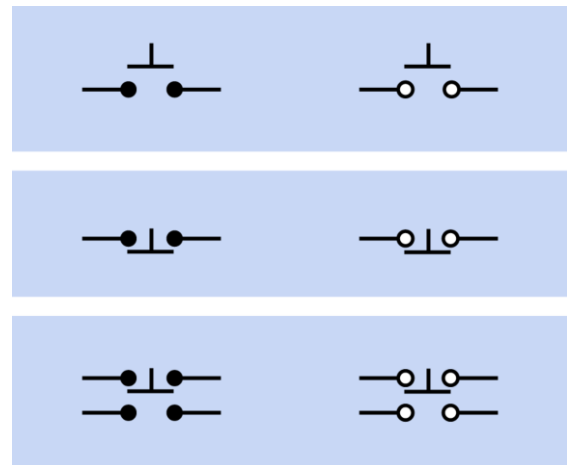


Figure 5-1. Commonly used schematic symbols to represent a simple pushbutton. See text for details.

How It Works

[Figure 5-4](#) shows a cross-section of a pushbutton that has a single steel return spring, to create resistance to downward force on the button, and a pair of springs above a pair of contacts, to hold each contact in place and make a firm connection when the button is pressed. The two upper contacts are electrically linked, although this feature is not shown.

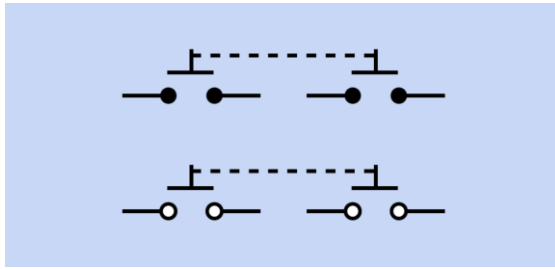


Figure 5-2. Commonly used schematic symbols to represent a double-pole pushbutton.

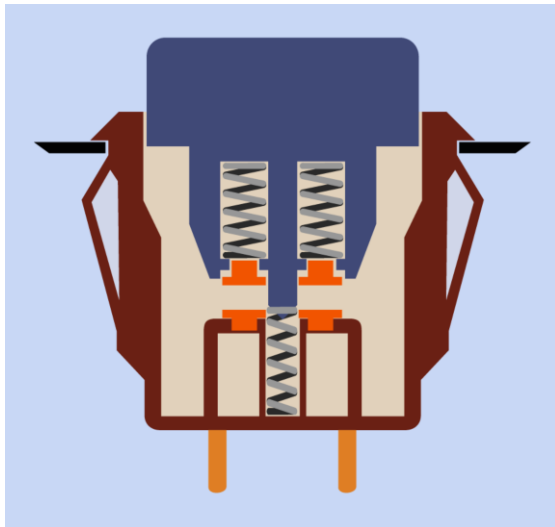


Figure 5-4. Cross-section of a pushbutton showing two spring-loaded contacts and a single return spring.

Variants

Poles and Throws

Abbreviations that identify the number of poles and contacts inside a pushbutton are the same as the abbreviations that identify those attributes in a **switch**. A few examples will make this clear:

SPST, also known as 1P1T

Single pole, single throw

DPST also known as 2P1T

Double pole, single throw



Figure 5-3. The simplest, traditional form of pushbutton, in which pressing the button creates a connection between two contacts.

SPDT also known as 1P2T

Single pole, double throw

3PST also known as 3P1T

Three pole, single throw

While a switch may have an additional center position, pushbuttons generally do not.

On-Off Behavior

Parentheses are used to indicate the momentary state of the pushbutton while it is pressed. It will return to the other state by default.

OFF-(ON) or (ON)-OFF

Contacts are normally open by default, and are closed only while the button is pressed. This is sometimes described as a *make-to-make* connection, or as a *Form A* pushbutton.

ON-(OFF) or (OFF)-ON

Contacts are normally closed by default, and are open only while the button is pressed. This is sometimes described as a *make-to-break* connection, or as a *Form B* pushbutton.

ON-(ON) or (ON)-ON

This is a double-throw pushbutton in which one set of contacts is normally closed. When the button is pressed, the first set of contacts

is opened and the other set of contacts is closed, until the button is released. This is sometimes described as a *Form C* pushbutton.

For a single-throw pushbutton, the terms NC or NO may be used to describe it as *normally closed* or *normally open*.

Slider

This type, also known as a *slide* pushbutton, contains a thin bar or rod that slides in and out of a long, narrow enclosure. Contacts on the rod rub across secondary contacts inside the enclosure. Closely resembling a slider switch, it is cheap, compact, and well adapted for multiple connections (up to 8 separate poles in some models). However, it can only tolerate low currents, has limited durability, and is vulnerable to contamination.

A four-pole, double-throw pushbutton is shown in Figure 5-5. A variety of plastic caps can be obtained to press-fit onto the end of the white nylon actuator.

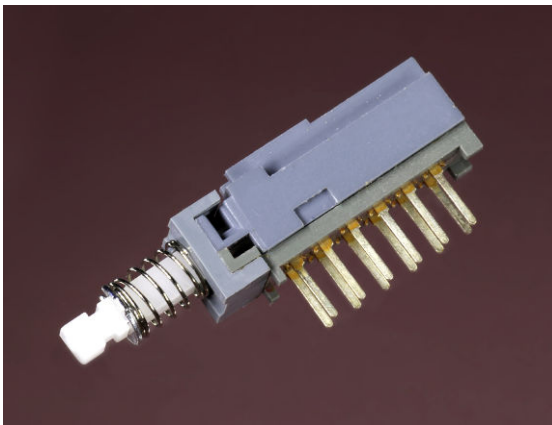


Figure 5-5. A 4PDT slider pushbutton, shown without the cap that can be snapped onto the end of the actuator.

Figure 5-6 shows schematic symbols for two possible slide pushbuttons, with a black rectangle indicating each sliding contact. The lead that functions as a pole is marked with a P in each case. Standardization for slide pushbutton sche-

matic symbols does not really exist, but these examples are fairly typical. An insulating section that connects the sliding contacts internally is shown here as a gray rectangle, but in some datasheets may appear as a line or an open rectangle.

Since the symbols for a slide pushbutton may be identical to the symbols for a slide switch, care must be taken when examining a schematic, to determine which type of component is intended.

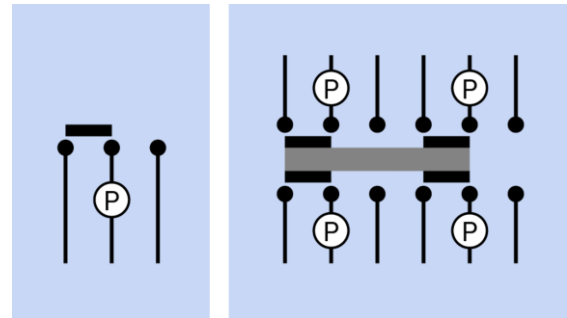


Figure 5-6. Left: schematic symbol for a simple SPDT slide pushbutton, where a movable contact shorts together either the left pair or right pair of fixed contacts. Right: A 4PDT pushbutton in which the same principle has been extended. The movable contacts are attached to each other mechanically by an insulator. Each pole terminal is marked with a P.

Styles

Many pushbutton switches are sold without caps attached. This allows the user to choose from a selection of styles and colors. Typically the cap is a push-fit onto the end of the rod or bar that activates the internal contacts. Some sample caps are shown in Figure 5-7, alongside a DPDT pushbutton. Any of the caps will snap-fit onto its actuator.

An *illuminated* pushbutton contains a small *incandescent bulb*, *neon bulb*, or **LED** (*light-emitting diode*). The light source almost always has its own two terminals, which are isolated from the other terminals on the button housing and can be wired to activate the light when the button is pressed, when it is released, or on some other basis. Pushbuttons containing LEDs usually



Figure 5-7. Caps (buttons or knobs) that may be sold as separate accessories for some pushbuttons, shown here alongside a compatible pushbutton switch.

require external series resistors, which should be chosen according to the voltage that will be used. See the **LED** entry in Volume 2 for additional commentary on appropriate series resistors. An example of an illuminated pushbutton is shown in [Figure 5-8](#). This is a DPDT component, designed to be mounted on a printed circuit board, with an additional lead at each end connecting with an internal LED underneath the translucent white button.

Termination and Contact Plating

These options are the same as for a **switch** and are described in that entry.

Mounting Style

The traditional panel-mounted button is usually secured through a hole in the panel by tightening a nut that engages with a thread on the bushing of the pushbutton. Alternatively, a push-

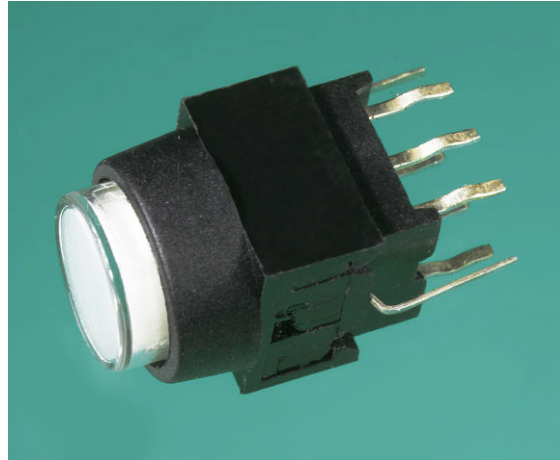


Figure 5-8. This pushbutton contains an LED underneath the white translucent button.

button housing can have flexible plastic protrusions on either side, allowing it to be snapped into place in an appropriate-sized panel cutout. This style is shown in [Figure 5-4](#).

PC pushbuttons (pushbuttons mounted in a printed circuit board, or PCB) are a common variant. After the component has been installed in the circuit board, either the button must align with a cutout in the front panel and poke through it when the device is assembled, or an external (non-electrical) button that is part of the product enclosure must press on the actuator of the pushbutton after assembly.

Surface-mount pushbuttons that allow direct fingertip access are uncommon. However, about one-quarter of tactile switches are designed for surface mount at the time of writing. They are typically found beneath membranes that the user presses to activate the switch beneath—for example, in remotes that are used to operate electronic devices.

Sealed or Unsealed

A sealed pushbutton will include protection against water, dust, dirt, and other environmental hazards, at some additional cost.

Latching

This variant, also known as a *press-twice* pushbutton, contains a mechanical ratchet, which is rotated each time the button is pressed. The first press causes contacts to latch in the closed state. The second press returns the contacts to the open state, after which, the process repeats. This press-twice design is typically found on flashlights, audio equipment, and in automotive applications. While *latching* is the most commonly used term, it is also known as *push-push*, *locking*, *push-lock push-release*, *push-on push-off*, and *alternate*.

In a latching pushbutton with *lockdown*, the button is visibly lower in the latched state than in the unlatched state. However, buttons that behave this way are not always identified as doing so on their datasheets.

A six-pole double-throw pushbutton that latches and then unlatches each time it is pressed is shown in Figure 5-9.

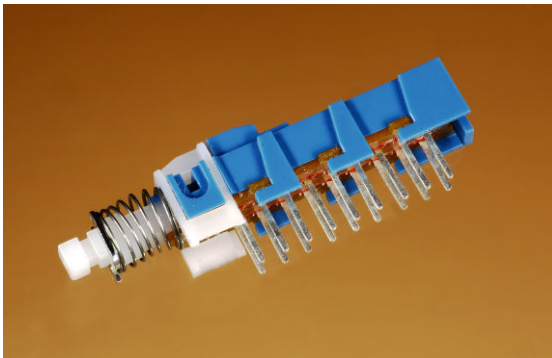


Figure 5-9. This 6PDT pushbutton latches and then unlatches, each time it is pressed.

Two more variants are shown in Figure 5-10. On the right is a simple DPDT latching pushbutton with lockdown. On the left is a latching pushbutton that cycles through four states, beginning with one “off” state, the remaining three connecting a different pair of its wires in turn.

A simple OFF-(ON) button may appear to have a latching output if it sends a pulse to a **microcontroller** in which software inside the micro-

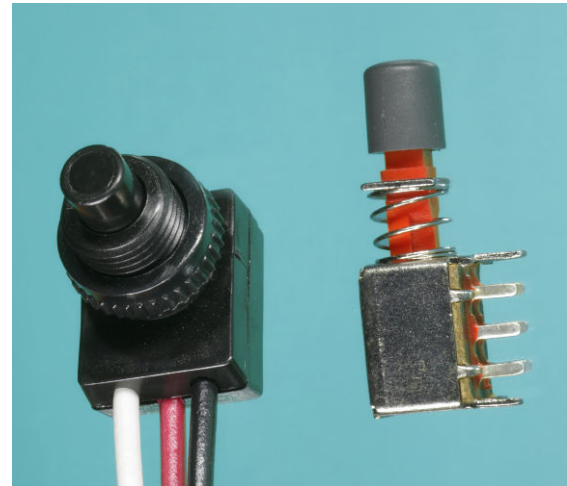


Figure 5-10. At right, a simple DPDT latching pushbutton with lockdown. At left, this pushbutton cycles through four states, one of them an “off” state, the others connecting a different pair of its wires in turn.

controller toggles an output between two states. The microcontroller can step through an unlimited number of options in response to each button press. Examples are found on cellular phones or portable media players.

A mechanically latching pushbutton has a higher failure rate than a simple OFF-(ON) button, as a result of its internal mechanism, but has the advantage of requiring no additional microcontroller to create its output. Microcontrollers are discussed in Volume 2.

Foot Pedal

Foot pedal pushbuttons generally require more actuation force than those intended for manual use. They are ruggedly built and are commonly found in vacuum cleaners, audio-transcription foot pedals, and “stomp boxes” used by musicians.

Keypad

A keypad is a rectangular array of usually 12 or 16 OFF-(ON) buttons. Their contacts are accessed via a *header* suitable for connection with a ribbon cable or insertion into a printed circuit board. In some keypads, each button connects with a

separate contact in the header, while all the buttons share a common ground. More often, the buttons are *matrix encoded*, meaning that each of them bridges a unique pair of conductors in a matrix. A 16-button matrix is shown in [Figure 5-11](#). This configuration is suitable for polling by a **microcontroller**, which can be programmed to send an output pulse to each of the four horizontal wires in turn. During each pulse, it checks the remaining four vertical wires in sequence, to determine which one, if any, is carrying a signal. Pullup or pulldown resistors should be added to the input wires to prevent the inputs of the microcontroller from behaving unpredictably when no signal is present. The external appearance of two keypads is shown in [Figure 5-12](#).

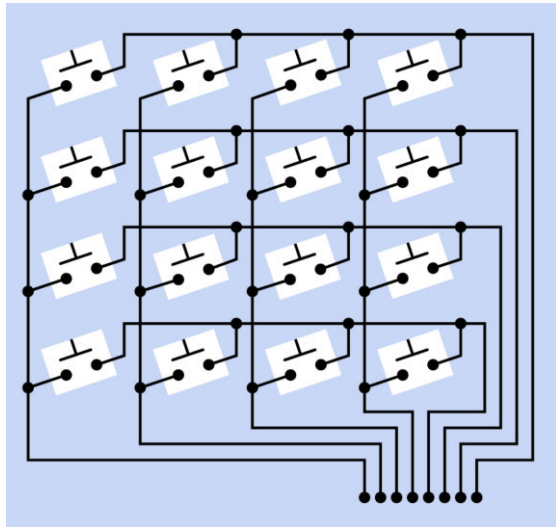


Figure 5-11. Buttons in a numeric keypad are usually wired as a matrix, where each button makes a connection between a unique pair of wires. This system is suitable for being polled by a microcontroller.

Tactile Switch

Despite being called a switch, this is a miniature pushbutton, less than 0.4" square, designed for insertion in a printed-circuit board or in a solderless breadboard. It is almost always a SPST device but may have four pins, one pair connected to each contact. Tactile switches may be PC-mounted behind membrane pads. An example is shown in [Figure 5-13](#).

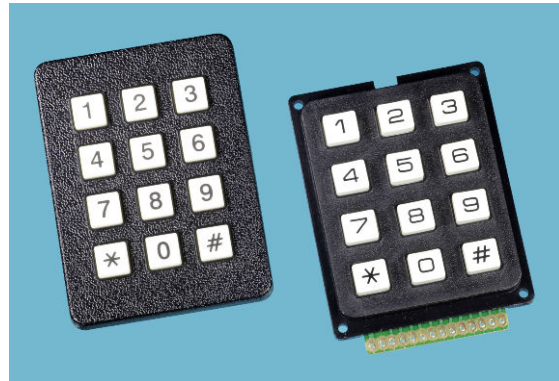


Figure 5-12. The keypad on the left is matrix-encoded, and is polled via seven through-hole pins that protrude behind it. The keypad on the right assigns each button to a separate contact in its header. See the text for details about matrix encoding.



Figure 5-13. A typical tactile switch.

Membrane Pad

Typically found on devices such as microwave ovens where contacts must be sealed against particles and liquids. Finger pressure on a membrane pad closes hidden or internal pushbuttons. They are usually custom-designed for specific product applications and are not generally available as generic off-the-shelf components. Some surplus pads may be found for sale on auction websites.

Radio Buttons

The term *radio buttons* is sometimes used to identify a set of pushbuttons that are mechanically interlinked so that only one of them can make an electrical connection at a time. If one button is pressed, it latches. If a second button is pressed, it latches while unlatching the first button. The buttons can be pressed in any sequence. This system is useful for applications such as component selection in a stereo system, where only one input can be permitted at a time. However, its use is becoming less common.

Snap-Action Switches

A snap-action switch (described in detail in the **switch** section of this encyclopedia) can be fitted with a pushbutton, as shown in [Figure 5-14](#). This provides a pleasingly precise action, high reliability, and capability of switching currents of around 5A. However, snap-action switches are almost always single-pole devices.

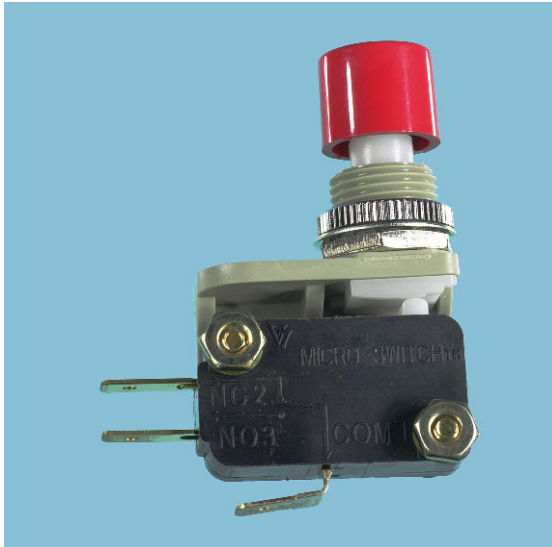


Figure 5-14. A pushbutton mounted on top of a SPDT snap-action switch.

Emergency Switch

An emergency switch is a normally-closed device, usually consisting of a large pushbutton

that clicks firmly into its “off” position when pressed, and does not spring back. A flange around the button allows it to be grasped and pulled outward to restore it to its “on” position.

Values

Pushbutton current ratings range from a few mA to 20A or more. Many pushbuttons have their current ratings printed on them but some do not. Current ratings are usually specified for a particular voltage, and may differ for AC versus DC.

How to Use it

Issues such as appearance, tactile feel, physical size, and ease of product assembly tend to dictate the choice of a pushbutton, after the fundamental requirements of voltage, current, and durability have been satisfied. Like any electromechanical component, a pushbutton is vulnerable to dirt and moisture. The ways in which a device may be used or abused should be taken into account when deciding whether the extra expense of a sealed component is justified.

When a pushbutton controls a device that has a high inductive load, a *snubber* can be added to minimize arcing. See “[Arcing](#)” ([page 47](#)) in the **switch** entry of this encyclopedia, for additional information.

What Can Go Wrong

No Button

When ordering a pushbutton switch, read data-sheets carefully to determine whether a cap is included. Caps are often sold separately and may not be interchangeable between switches from different manufacturers.

Mounting Problems

In a panel-mount pushbutton that is secured by turning a nut, the nut may loosen with use, allowing the component to fall inside its enclosure when the button is pressed. Conversely, over-tightening the nut may strip the threads on the

pushbutton bushing, especially in cheaper components where the threads are molded into plastic. Consider applying a drop of Loc-Tite or similar adhesive before completely tightening the nut. Nut sizes vary widely, and finding a replacement may be time-consuming.

LED Issues

When using a pushbutton containing an LED, be careful to distinguish the LED power terminals from the switched terminals. The manufacturer's datasheet should clarify this distinction, but the polarity of the LED terminals may not be clearly indicated. If a diode-testing meter function is un-

available, a sample of the switch should be tested with a source of 3 to 5VDC and a 2K series resistor. Briefly touching the power to the LED terminals, through the resistor, should cause the LED to flash dimly if the polarity is correct, but should not be sufficient to burn out the LED if the polarity is incorrect.

Other Problems

Problems such as arcing, overload, short circuits, wrong terminal type, and contact bounce are generally the same as those associated with a **switch**, and are summarized in that entry in this encyclopedia.

switch



The term **switch** refers here to a physically operated mechanical switch, controlled by flipping a lever or sliding a knob. Although there is some overlap of function, **rotary switches** and **pushbuttons** have their own separate entries. Solid-state switching components are described in entries for **bipolar transistor**, **unijunction transistor**, and **field-effect transistor**. Integrated-circuit switching devices will be found in Volume 2. *Coaxial switches* are used for high-frequency signals, and are not included in this encyclopedia. *Multidirectional switches* differentiate up, down, left, right, diagonal, rotational, and other finger inputs, and are not included in this encyclopedia.

OTHER RELATED COMPONENTS

- **pushbutton** (See [Chapter 5](#))
- **rotary switch** (See [Chapter 7](#))

What It Does

A switch contains at least two contacts, which close or open when an external lever or knob is flipped or moved. Schematic symbols for the most basic type of on-off switch are shown in [Figure 6-1](#).

The most fundamental type of switch is a *knife switch*, illustrated in [Figure 6-2](#). Although it was common in the earliest days of electrical discovery, today it is restricted to educational purposes in schools, and (in a more robust format) to AC electrical supply panels, where the large contact area makes it appropriate for conducting high amperages, and it can be used for “hot switching” a substantial load.

How It Works

The *pole* of a switch is generally connected with a movable contact that makes or breaks a connection with a secondary contact. If there is only one pole, this is a *single-pole* switch. If there is an additional pole, electrically isolated from the

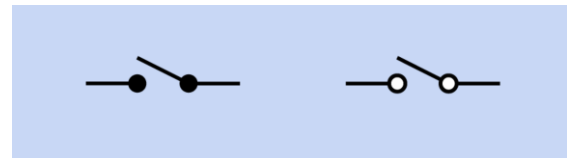


Figure 6-1. The two most common schematic symbols for a SPST switch, also known as an on-off switch. The symbols are functionally identical.

first, with its own contact or set of contacts, this is a *two-pole* switch, also known as a *double-pole* switch. Switches with more than 4 poles are uncommon.

If there is only one secondary contact per pole, this is a *single-throw* or *ST* switch, which may also be described as an *on-off* or *off-on* switch. If there is an additional secondary contact per pole, and the pole of the switch connects with the second contact while disconnecting from the first, this is a *double-throw* or *DT* switch, also known as a *two-way* switch.



Figure 6-2. A DPST knife switch intended for the educational environment.

A double-throw switch may have an additional center position. This position may have no connection (it is an “off” position) or in some cases it connects with a third contact.

Where a switch is spring-loaded to return to one of its positions when manual pressure is released, it functions like a **pushbutton** even though its physical appearance may be indistinguishable from a switch.

Variants

Float switch, mercury switch, reed switch, pressure switch, and Hall-effect switch are considered as sensing devices, and will be found in Volume 3.

Terminology

Many different types of switches contain parts that serve the same common functions. The *actuator* is the lever, knob, or toggle that the user turns or pushes. A *bushing* surrounds the actuator on a toggle-type switch. The *common contact* inside a switch is connected with the *pole* of

the switch. Usually a *movable contact* is attached to it internally, to touch the secondary contact, also known as a *stationary contact* when the movable contact is flipped to and fro.

Poles and Throws

Abbreviations identify the number of poles and contacts inside a switch. A few examples will make this clear:

SPST also known as 1P1T

Single pole, single throw

DPST also known as 2P1T

Double pole, single throw

SPDT also known as 1P2T

Single pole, double throw

3PST also known as 3P1T

Three pole, single throw

Other combinations are possible.

In [Figure 6-3](#), schematic symbols are shown for double-throw switches with 1, 2, and 3 poles. The dashed lines indicate a mechanical connection, so that all sections of the switch move together when the switch is turned. No electrical connection exists between the poles.

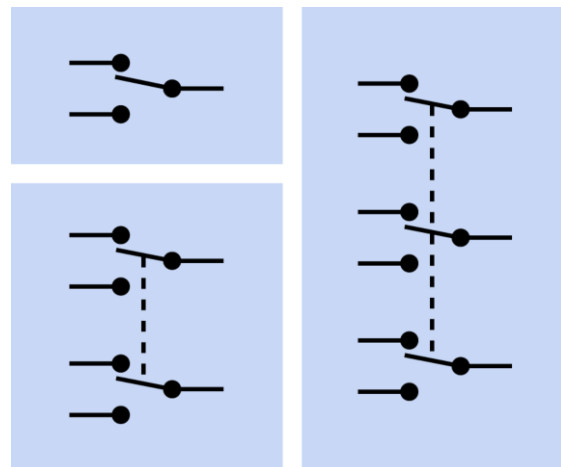


Figure 6-3. Schematic symbols to represent three types of double-throw switch. Top left: Single-pole. Bottom left: Double-pole. Right: Triple-pole, more commonly known as 3-pole.

On-Off Behavior

The words ON and OFF are used to indicate the possible states of a switch. The additional word NONE is used by some manufacturers to indicate that a switch does not have a center position. Some manufacturers don't bother with the word NONE, assuming that if the word is omitted, a center position does not exist.

ON-OFF or ON-NONE-OFF

A basic on-off SPST switch with no center position.

ON-ON or ON-NONE-ON

A basic SPDT switch with no center position.

ON-OFF-ON

A double-throw switch with center-off position (no connection when the switch is centered).

ON-ON-ON

A triple-throw switch where the center position connects with its own set of terminals.

Parentheses are used in descriptions of spring-loaded switches to indicate a momentary state that lasts only as long as pressure is applied to the actuator.

(ON)-OFF or OFF-(ON)

A spring-loaded switch that is normally off and returns to that position when pressure is released. Also known as **NO** (normally open), and sometimes described as **FORM A**. Its performance is similar to that of a push-button and is sometimes described as a **make-to-make** connection.

ON-(OFF) or (OFF)-ON

A spring-loaded switch that is normally on and returns to that position when pressure is released. This is sometimes described as a **make-to-break** connection. Also known as **NC** (normally closed), and sometimes described as **FORM B**.

(ON)-OFF-(ON)

A spring-loaded double-throw switch with a no-connection center position to which it returns when pressure on its actuator is released.

Other combinations of these terms are possible.

Most double-throw switches break the connection with one contact (or set of contacts) before making the connection with the second contact (or set of contacts). This is known as a **break before make** switch. Much less common is a **make before break** switch, also known as a **shorting switch**, which establishes the second connection a moment before the first connection is broken. Use of a shorting switch may cause unforeseen consequences in electronic components attached to it, as both sides of the switch will be briefly connected when the switch is turned.

Snap-Action

Also known as a **limit switch** and sometimes as a **microswitch** or **basic switch**. This utilitarian design is often intended to be triggered mechanically rather than with finger pressure, for example in 3D printers. It is generally cheap but reliable.

Two snap-action switches are shown in [Figure 6-4](#). A sectional view of a snap-action ON-(ON) limit switch is shown in [Figure 6-5](#). The pole contacts are mounted on a flexible strip which can move up and down in the center of the switch. The strip has a cutout which allows an inverted U-shaped spring to flip to and fro. It keeps the contacts pressed together in either of the switch states.

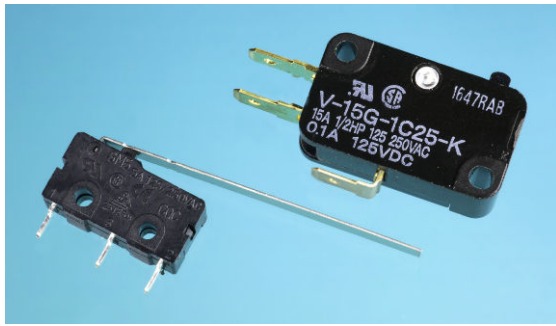


Figure 6-4. Two SPDT snap-action switches, also known as limit switches. The one on the right is full-size. The one on the left is miniature, with an actuator arm to provide additional leverage. The arm may be trimmed to the required length.

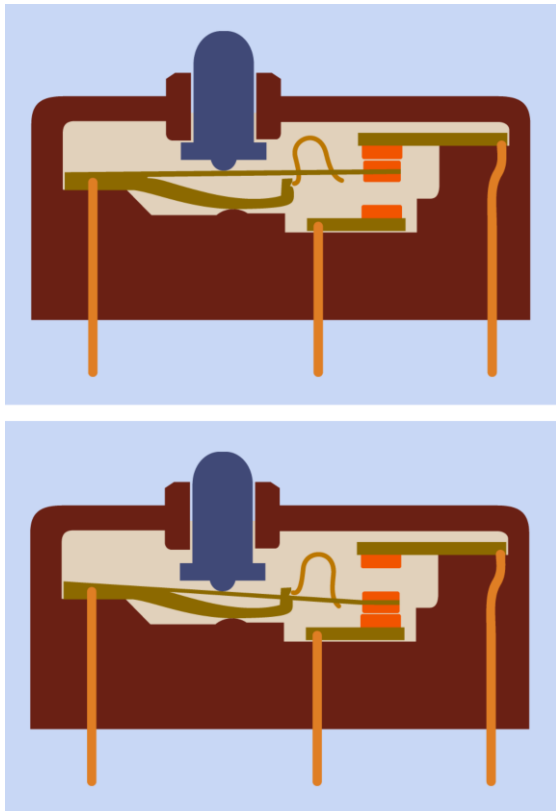


Figure 6-5. Top: Two contacts inside this limit switch are touching by default. Bottom: When the external button is pressed, it pushes a flexible metal strip downward until it connects with the lower contact. The inverted-U-shaped component is a spring that rests inside a cutout in the flexible strip and resists motion through the central part of its travel.

The term *snap action* refers to a spring-loaded internal mechanism which snaps to and fro between its two positions. This type of switch is usually SPDT and has a momentary action; in other words, it functions in ON-(ON) mode, although OFF-(ON) and (less often) ON-(OFF) versions are available. The body of the switch is sealed, with a small button protruding through a hole. A thin metal arm may provide additional leverage to press the button. A roller may be mounted at the end of the arm so that the switch can be activated as it slides against a moving mechanical component such as a cam or a wheel. The switch is commonly used to limit the travel or rotation of such a component. Literally thousands of variants are available, in different sizes, requiring different amounts of force for activation. Subminiature snap-action switches can often be actuated by a pressure of only a few grams.

Rocker

Three rocker switches are shown in Figure 6-6. A sectional view of a rocker switch is shown in Figure 6-7. In this design, a spring-loaded ball bearing rolls to either end of a central rocker arm when the switch is turned. Rocker switches are often used as power on-off switches.

Slider

Many types of slider switch (also known as *slide switch*) are widely used as a low-cost but versatile way to control small electronic devices, from clock-radios to stereos. The switch is usually mounted on a circuit board, and its knob or cap protrudes through a slot in the panel. This design is more vulnerable to dirt and moisture than other types of switch. It is usually cheaper than a toggle switch but is seldom designed for use with a high current.

Most slide switches have two positions, and function as SPDT or DPDT switches, but other configurations are less commonly available with more poles and/or positions. A subminiature slide switch is shown in Figure 6-8, while some schematic representations are shown in Figure 6-9,

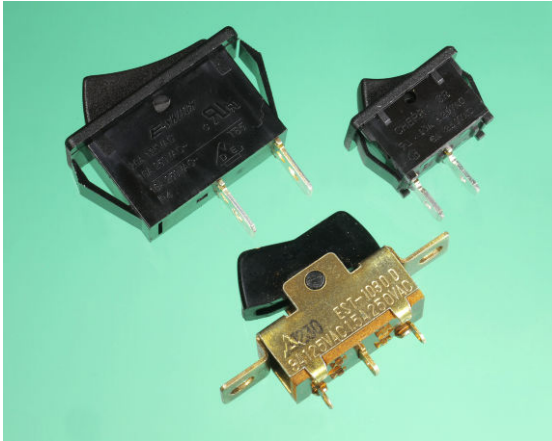


Figure 6-6. Three rocker switches, the upper two designed for push-insertion into a suitably sized rectangular hole in a panel. The switch at front-center is intended to be screwed in place, and is more than 20 years old, showing that while the choice of materials has changed, the basic design has not.

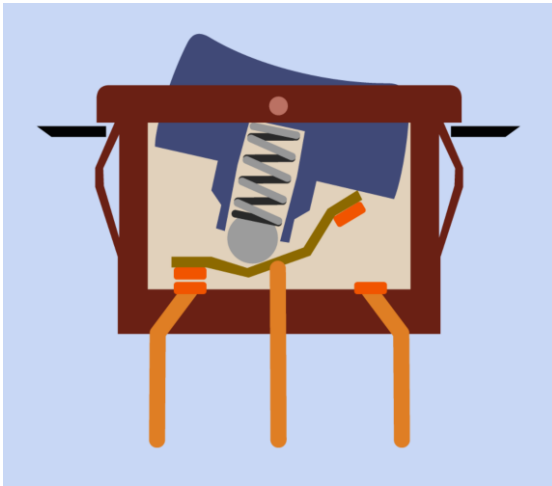


Figure 6-7. This sectional view of a rocker switch shows a spring-loaded ball-bearing that rolls to and fro along a rocker arm, connecting either pair of contacts when the switch is turned.

where a black rectangle indicates a sliding internal contact, and a terminal that functions as a pole is identified with letter P in each case. Top left: A SPDT switch using a two-position slider. Top right: A 4PDT slide switch. Bottom left: There are no poles in this switch, as such. The slider can

short together any of four pairs of contacts. Bottom right: The slider shorts together three possible pairs of contacts out of four. Here again, there is no pole.

Note that the schematic representation of a slide switch may be identical to that of a slide **push-button**. A schematic should be inspected carefully to determine which type is intended.

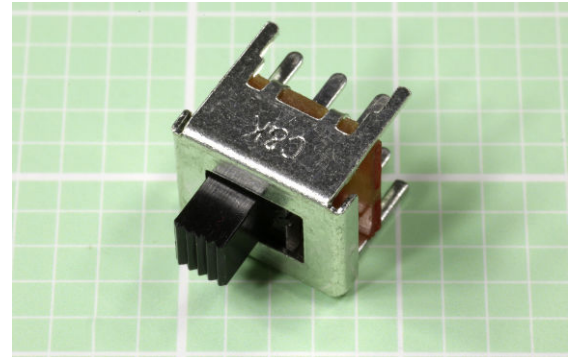


Figure 6-8. This subminiature slide switch is less than half an inch long, rated 0.3A at 30VDC. Larger versions look almost identical, but can handle only slightly more current.

The representation of sliders in schematics has not been standardized, but the samples shown are common.

Toggle

A *toggle switch* provides a firm and precise action via a lever (the *toggle*) that is usually tear-drop shaped and nickel plated, although plastic toggles are common in cheaper variants. Formerly used to control almost all electronic components (including early computers), the toggle has declined in popularity but is still used in applications such as automobile accessory kits, motorboat instrument panels, and industrial controls.

Three miniature DPDT toggle switches are shown in [Figure 6-10](#). Two full-size, heavy-duty toggle switches are shown in [Figure 6-11](#). A full-size, four-pole, double-throw heavy-duty toggle switch is shown in [Figure 6-12](#). Toggle switches with more poles are extremely rare.

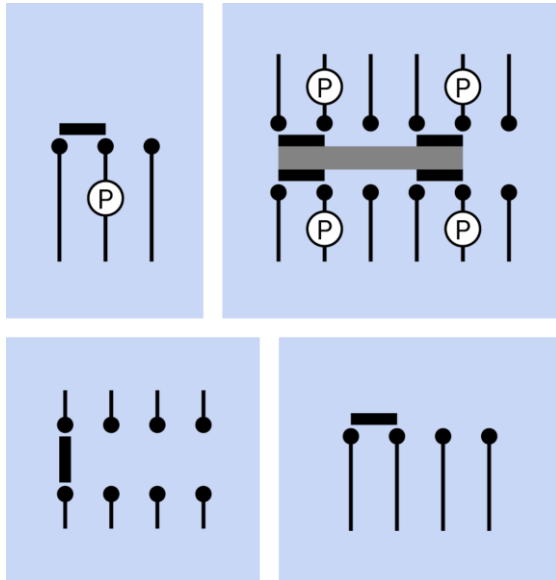


Figure 6-9. Slide switch schematics. Each black rectangle represents a movable contact that connects two pairs of fixed contacts at a time. Detailed commentary on these variants will be found in the body of the text. Manufacturers may use variants of these symbols in their datasheets (for example, the gray rectangle indicating an insulating contact carrier, at top right, may be represented as a single line, or a black outline with a white center).

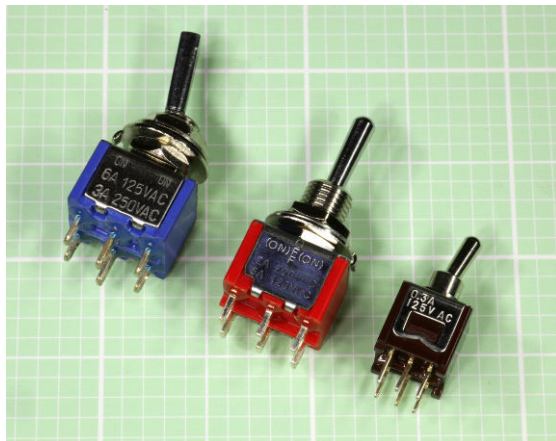


Figure 6-10. Three miniature toggle switches with current ratings ranging from 0.3A to 6A at 125VAC. Each small square in the background grid measures 0.1" x 0.1".

An automotive toggle switch is shown in [Figure 6-13](#). Its plastic toggle is extended to minimize operating error.



Figure 6-11. Two full-size toggle switches capable of handling significant current. At left, the switch terminates in quick-connect terminals. At right, the switch has solder terminals (some of them containing residual traces of solder).

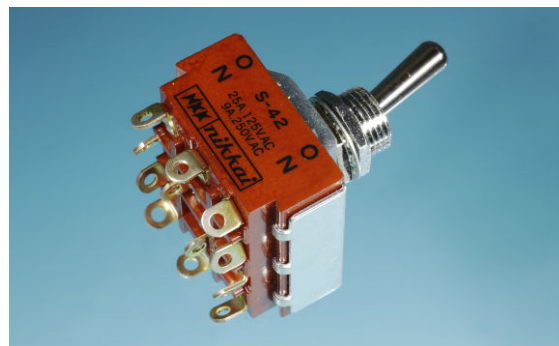


Figure 6-12. A 4PDT full-size toggle switch with solder terminals, capable of switching 25A at 125VAC. Four-pole switches are relatively unusual.



Figure 6-13. A toggle switch intended for control of automotive accessories.

High-end toggle switches are extremely durable and can be sealed from environmental

contamination with a thin *boot* made from molded rubber or vinyl, which screws in place over the toggle, using the thread on the switch bushing. See [Figure 6-14](#).



Figure 6-14. A rubber or vinyl boot can be used to protect a toggle switch from contamination with dirt or water. Each boot contains a nut that screws onto the threads of a toggle switch, as shown at left.

A *locking toggle switch* has a toggle that must be pulled out against the force of a retaining spring, before the toggle can be moved from one position to another. The toggle then snaps back into place, usually engaging in a small slot in the bushing of the switch.

DIP

A DIP switch is an array of very small, separate switches, designed for mounting directly on a circuit board, either in through-hole or surface-mount format. Through-hole DIP switches have two rows of pins with a 0.1" pitch, the rows being spaced 0.3" apart to fit a standard DIP (dual-inline package) socket or comparable configuration of holes in the board. Surface-mount DIP switches may have 0.1" or 0.05" pitch.

Most DIP arrays consist of SPST switches, each of which can close or open a connection between two pins on opposite sides of the switch body. The switch positions are usually labelled ON and OFF. [Figure 6-15](#) shows a selection of DIP switches. [Figure 6-16](#) shows the internal connections in a DIP switch.

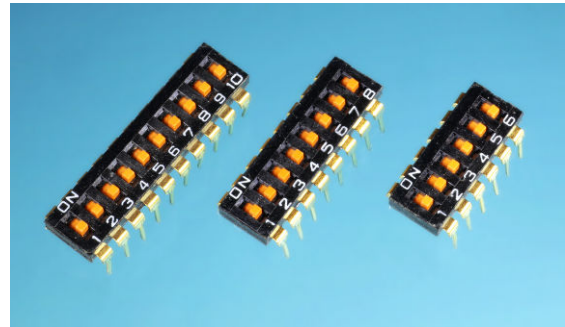


Figure 6-15. As shown here, DIP switches are available with a variety of "positions," meaning the number of switches, not the number of switch states.

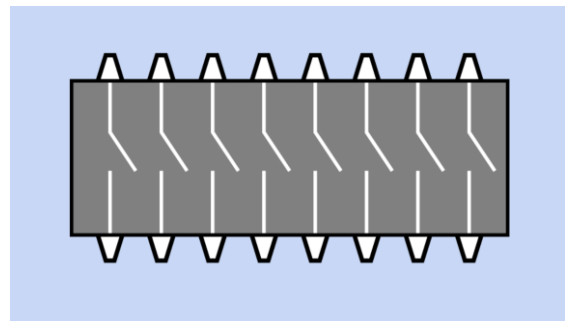


Figure 6-16. The interior connections of a 16-pin DIP switch.

The number of switches in a DIP array is usually referred to as its number of "positions." This should not be confused with the two positions of each physical switch lever. SPST DIP switches are made with 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, and 16 positions.

Early IBM-compatible desktop computers often required the user to set the position of an internal DIP switch when making routine upgrades such as installing an additional disk drive. While this feature is now obsolete, DIP switches are still used in scientific equipment where the user is expected to be sufficiently competent to open a cabinet and poke around inside it. Because of the 0.1" spacing, a small screwdriver or the tip of a pen is more appropriate than a finger to flip individual levers to and fro.

DIP switches may also be used during prototype development, as they allow a convenient way to test a circuit in numerous different modes of operation.

Most DIP switches have wire terminals which are just long enough for insertion into a standard breadboard.

DIP switch package options include standard, low-profile, right-angle (standing at 90 degrees relative to the circuit board), and piano (with switch levers designed to be pressed, like tiny rocker switches, instead of being flipped to and fro).

Some SPDT, DPST, DPDT, 3PST, and 4PST variants exist, but are uncommon. Multiple external pins connect with the additional internal switch contacts, and a manufacturer's datasheet should be consulted to confirm the pattern of internal connections. A surface-mount, 0.1" pitch, DPST DIP switch is shown in Figure 6-17, with a plastic cover to protect the switches from contamination during wave soldering (at left), and with the cover peeled off (at right).

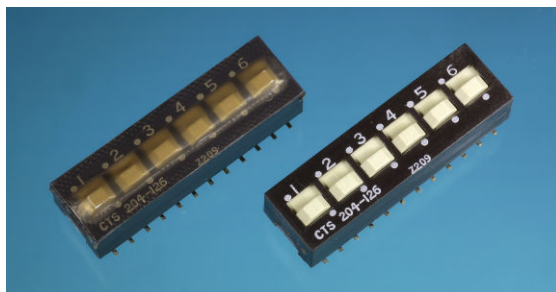


Figure 6-17. A SPDT surface-mount double-throw DIP switch, sold with a plastic cover (shown at left) to protect it during wave soldering. The cover has been removed at right.

SIP

A SIP switch is an array of small, separate switches, identical in concept to a DIP switch, but using only one row of pins instead of a double row. The applications for SIP switches are the

same as DIP switches; the primary difference is simply that the SIP switch occupies a little less space, while being perhaps slightly less convenient to use.

One terminal of each switch usually shares a common bus. The internal connections in a typical 8-pin SIP array are shown in Figure 6-18. Pin spacing is 0.1", as in a typical DIP switch.

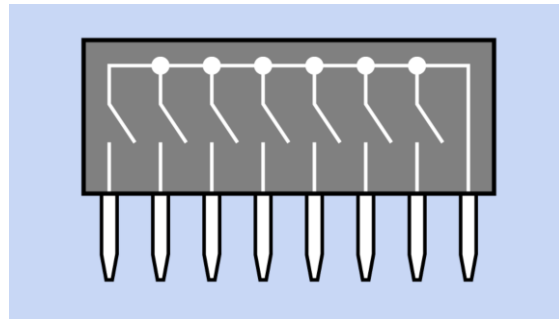


Figure 6-18. The interior connections of an 8-pin SIP switch incorporating a common bus.

Paddle

A paddle switch has a flat-sided tab-shaped plastic actuator, relatively large to allow a firm, error-free grip. Internally it is often comparable with a rocker switch, and is generally used with AC power. Some toggle-switch bodies are also sold with paddle-shaped actuators. A subminiature paddle switch is shown in Figure 6-19.

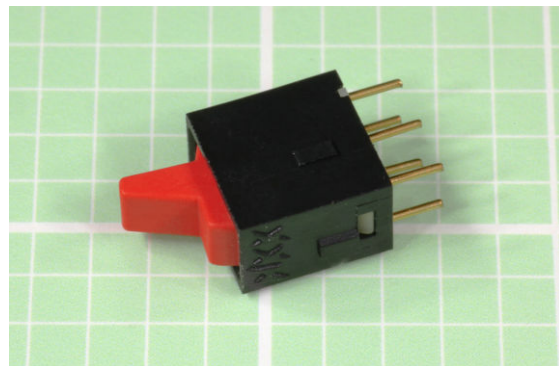


Figure 6-19. A subminiature paddle switch. Full-size versions are often used as power switches.

Vandal Resistant Switch

Typically fabricated from stainless steel, this is designed to withstand most types of abuse and is also weather-proof. The pushbuttons that allow pedestrians to trigger a traffic signal are a form of vandal-resistant switch.

Tactile Switch

This is considered to be a **pushbutton**, and is described in that entry. See “[Tactile Switch](#)” (page 34).

Mounting Options

A *panel mount* switch generally has a threaded bushing that is inserted from behind the front panel of a product, through a hole of appropriate size. It is supplied with a lockwasher and a nut (often, two nuts) that fit the thread on the switch bushing.

Front panel mount usually means that screws visible from the front of the panel are attached to a bracket on the switch behind the panel. The actuator of the switch is accessible through a cutout in the panel. This mounting style is mostly used for rocker switches and sometimes for slide switches.

Subpanel mount means that the switch is attached to a separate plate or chassis behind the control panel. The actuator of the switch is accessible through a cutout.

Snap-in mount requires a switch with flexible plastic or metal tabs each side, designed to push through a cutout in the panel, at which point the tabs spring out and retain the switch.

PC mount switches have pins that are soldered into a printed circuit board. They may have additional solderable lugs to provide mechanical support.

Surface mount switches are attached to a board in the same manner as other surface-mount components.

Termination

Switches (and pushbuttons) are available with a variety of terminals.

Solder lugs are small tabs, each usually perforated with a hole through which the end of a wire can be inserted prior to soldering.

PC terminals are pins that protrude from the bottom of the switch, suitable for insertion in a printed circuit board. This style is also known as a *through-hole*. The terminals may have a right-angle bend to allow the component to be mounted flat against the board, with the switch actuator sticking out at the side. This termination style is known as *right-angle PC*. Many manufacturers offer a choice of straight or bent pin terminals, but the component may be listed in a catalog under either of those options, with no indication that other options exist. Check manufacturer datasheets carefully.

Quick connect terminals are spade-shaped to accept push-on connectors, commonly used in automotive applications. Hybrid quick-connect terminals that can also function as solder lugs are sometimes offered as an option.

Screw terminals have screws premounted in flat terminals, for solderless attachment of wires.

Wire leads are flexible insulated wires, often with stripped and tinned ends, protruding at least an inch from the body of the component. This option is becoming uncommon.

Contact Plating Options

The internal electrical contacts of a switch are usually plated with silver or gold. Nickel, tin, and silver alloys are cheaper but less common. Other types are relatively rare.

Values

Switches designed for electronic devices vary widely in power capability, depending on their purpose. Rocker switches, paddle switches, and toggle switches are often used to turn power on and off, and are typically rated for 10A at 125VAC,

although some toggle switches go as high as 30A. Snap-action or limit switches may be similarly rated, although miniature versions will have reduced capability. Slide switches cannot handle significant power, and are often rated around 0.5A (or less) at 30VDC. DIP and SIP switches have a typical maximum rating of 100mA at 50V and are not designed for frequent use. Generally they are used only when the power to the device is off.

How to Use it

Power Switches

When a simple SPST switch is used to turn DC power on and off, it conventionally switches the positive side of the power supply, also sometimes known as the *high side*. The primary reason for following this convention is that it is widely used; thus, following it will reduce confusion.

More importantly, an on-off switch that controls AC power must be used on the “live” side of the supply, not the “neutral” side. If you have any doubts about these concepts (which go beyond the scope of this book), consult a reference guide on this subject. Using a DPST component to switch both sides of an AC supply may be a worthwhile additional precaution in some applications. The ground wire of an AC supply should never be switched, because the device should always be grounded when it is plugged into an electric outlet.

Limit Switches

An application for two limit switches with a DC motor and two rectifier **diodes** is shown in [Figure 6-20](#). This diagram assumes that the motor turns clockwise when its lower terminal is positive, and counter-clockwise when its upper terminal is positive. Only two terminals are used (and shown) in each limit switch; they are chosen to be normally-closed. Other terminals inside a switch may exist, may be normally-open, and can be ignored.

The motor is driven through a dual-coil, DPDT latching **relay**, which will remain in either posi-

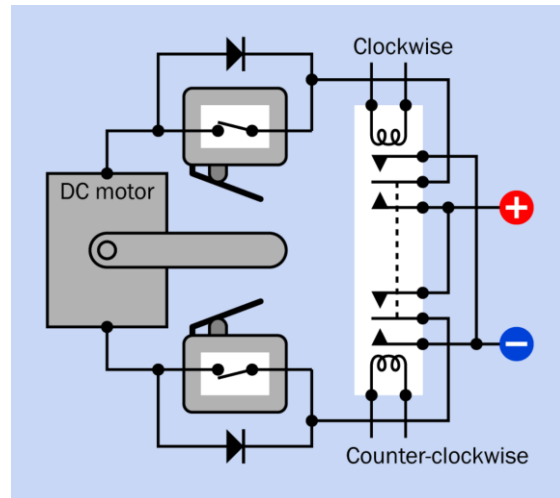


Figure 6-20. In this schematic, normally-closed limit switches are opened by pressure from an arm attached to a motor, thus switching off its power at each end of its permitted travel and preventing overload and burnout. A two-coil *latching relay* activates the motor. Rectifier diodes allow power to reach the motor to reverse its rotation when a limit switch is open.

tion indefinitely without drawing power. When the upper coil of the relay receives a pulse from a pushbutton or some other source, the relay flips to its upper position, which conducts positive current through the lower limit switch, to the lower terminal of the motor. The motor turns clockwise until the arm attached to its shaft hits the lower limit switch and opens it. Positive current is blocked by the lower diode, so the motor stops.

When the lower coil of the relay is activated, the relay flips to its lower position. Positive current can now reach the upper side of the motor through the upper limit switch. The motor runs counter-clockwise until its arm opens the upper limit switch, at which point the motor stops again. This simple system allows a DC motor to be run in either direction by a button-press of any duration, without risk of burnout when the motor reaches the end of its travel. It has been used for applications such as raising and lowering powered windows in an automobile.

A DPDT pushbutton could be substituted for the latching relay if manual control, only, is acceptable. However, in this scenario, sustained pressure on the pushbutton would be necessary to move the motor arm all the way to the opposite end of its travel. A DPDT switch might be more appropriate than a pushbutton.

Logic Circuits

Logic circuits that depend purely on switches can be constructed (for example, to add binary numbers) but are rare and have no practical applications. The most familiar and simplest example of manually switched logic is a pair of SPDT switches in house wiring, one positioned at the top of a flight of stairs and the other at the bottom, as shown in Figure 6-21. Either switch will turn the light on if it is currently off, or off if it is currently on. To extend this circuit by incorporating a third switch that has the same function as the other two, a DPDT switch must be inserted. See Figure 6-22.

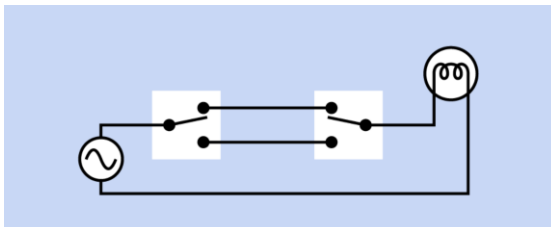


Figure 6-21. SPDT switches are commonly used in house wiring so that either of them will turn a shared light on if it is off, or off if it is on.

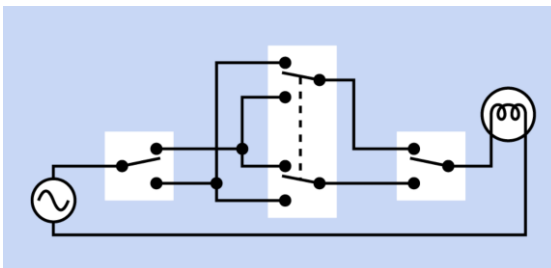


Figure 6-22. A DPDT switch must be inserted if three switches must have identical function to control the on-off state of a single light bulb.

Alternatives

As **microcontrollers** have become cheaper and more ubiquitous, they have taken over many functions in electronic products that used to be served by switches. A menuing system driven by a microcontroller can use one **rotational encoder** with a SPST pushbutton built into it to select and adjust numerous parameters in a device such as a car stereo, where functions were once selected and adjusted by individual switches and potentiometers. The rotational-encoder option takes up less space, is cheaper to build (assuming a microcontroller is going to be used in the device for other purposes anyway), and can be more reliable, as it reduces the number of electromechanical parts. Whether it is easier to use is a matter of taste. Cost and ergonomics may be the primary factors to consider when choosing where and how to use switches.

What Can Go Wrong

Arcing

The contacts inside a switch will be rapidly eroded if **arcing** (pronounced “arking”) occurs. An electric arc is a spark that tends to form when a switch is opened while conducting a high current or high voltage (typically 10A or more and 100V or more). The most common cause is an inductive load that generates back-EMF when it is switched on and forward-EMF when it is switched off. The surge can be many times the amperage that the load draws during continuous operation. In DC circuits, arcing can be reduced by using a rectifier **diode** in parallel with the load (with its polarity blocking normal current flow). This is often referred to as a **flyback diode** or **freewheeling diode**. In AC circuits, where a diode cannot be used in this way, a **snubber** (a simple combination of capacitor and resistor) may be placed around the load. A snubber can also be used around the switch itself, in DC circuits. See “Snubber” (page 108).

When switching an inductive load, it is generally prudent to use switches rated for twice as much current as the circuit will normally draw.

Dry Joints

Switches that control significant current will have substantial terminals, and these terminals will be attached to heavy-gauge wire. When using solder to make this type of connection, the combined heat capacity of the wire and the terminal will sink a lot more heat than a small component on a circuit board. At least a 30W soldering iron should be used. Lower-wattage irons may be incapable of melting the solder completely (even though they seem to), and a “dry joint” will result, which can have a relatively high electrical resistance and will be mechanically weak, liable to break later. Any good solder joint should withstand flexing of the wire attached to it.

Short Circuits

Because many switches are still wired in with solder tabs, screw terminals, or quick-connect terminals, wires that become accidentally detached can be a significant hazard. Heat-shrink tubing should be applied to enclose wires and terminals at the rear of a power switch, as an additional precaution. Power switches should always be used in conjunction with appropriate **fuses**.

Contact Contamination

Sealed switches should be used in any environment where dirt or water may be present. Slide switches are especially vulnerable to contamination, and are difficult to seal. Switches used in audio components will create “scratchy” sounds if their contacts deteriorate.

Wrong Terminal Type

Because switches are available with a wide variety of terminal types, it’s easy to order the wrong type. Switches may be supplied with pins for through-hole insertion in circuit boards; screw terminals; quick-disconnect terminals; or solder lugs. Variants may also be available for surface

mount. If a project requires, for example, the insertion of pins in a printed circuit board, and a switch is supplied with solder lugs, it will be unusable.

Part numbers generally include codes to identify each terminal variant, and should be studied carefully.

Contact Bounce

Also known as *switch bounce*. When two contacts snap together, extremely rapid, microscopic vibrations occur that cause brief interruptions before the contacts settle. While this phenomenon is not perceptible to human senses, it can be perceived as a series of multiple pulses by a **logic chip**. For this reason, various strategies are used to *debounce* a switch that drives a logic input. This issue is explored in detail in the entry on logic chips in Volume 2 of the encyclopedia.

Mechanical Wear

Any toggle or rocker switch contains a mechanical pivot, which tend to deteriorate in harsh environments. Friction is also an issue inside these switches, as the design often entails the rounded tip of a lever rubbing to and fro across the center of a movable contact.

The spring inside a snap switch or limit switch may fail as a result of metal fatigue, although this is rare. A slide switch is far less durable, as its contacts rub across each other every time the switch changes position.

In any application that entails frequent switching, or where switch failure is a critical issue, the most sensible practice is to avoid using cheap switches.

Mounting Problems

In a panel-mount switch that is secured by turning a nut, the nut may loosen with use, allowing the component to fall inside its enclosure. Conversely, overtightening the nut may strip the threads on the switch body, especially in cheaper components where the threads are molded into

plastic. Consider applying a drop of Loc-Tite or similar adhesive after moderately tightening the nut. Note that nut sizes vary widely, and finding a replacement may be time-consuming.

Cryptic Schematics

In some circuit schematics, the poles of a multi-pole switch may be visually separated from each

other, even at opposite sides of the page, for convenience in drawing the schematic. Dotted lines usually, but not always, link the poles. In the absence of dotted lines, switch segments are often coded to indicate their commonality. For example, SW1(a) and SW1(b) are almost certainly different parts of the same switch, with linked poles.

rotary switch

Not to be confused with **rotational encoder**, which has its own entry in this encyclopedia.

OTHER RELATED COMPONENTS

- **switch** (See [Chapter 6](#))
- **rotational encoder** (See [Chapter 8](#))

What It Does

A rotary switch makes an electrical connection between a *rotor*, mounted on a shaft that is turned by a knob, and one of two or more stationary contacts. Traditionally, it was the component of choice to select wavebands on a radio receiver, broadcast channels on a television or inputs on a stereo preamplifier. Since the 1990s, it has been substantially superseded by the **rotational encoder**. However it still has applications in military equipment, field equipment, industrial control systems, and other applications requiring a rugged component that will withstand heavy use and a possibly harsh environment. Also, while the output from a rotational encoder must be decoded and interpreted by a device such as a microcontroller, a rotary switch is an entirely passive component that does not require any additional electronics for its functionality.

Two typical schematic symbols for a rotary switch are shown in [Figure 7-1](#). They are functionally identical. A simplified rendering of the interior of a traditional-style rotary switch is shown in [Figure 7-2](#). A separate contact (not shown) connects with the rotor, which connects with each

of the stationary contacts in turn. The colors were chosen to differentiate the parts more clearly, and do not correspond with colors in an actual switch.

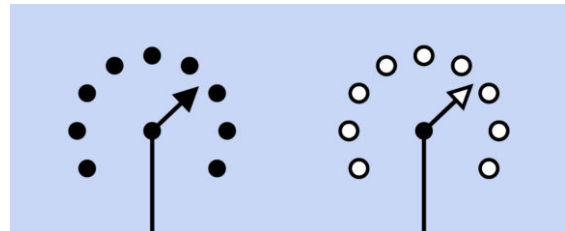


Figure 7-1. Typical schematic symbols for a rotary switch. The two symbols are functionally identical. The number of contacts will vary depending on the switch.

A selection of rotary switches is shown in [Figure 7-3](#). At top-left is an *open frame* switch, providing no protection to its contacts from contaminants. This type of component is now rare. At top-right is a twelve-position, single-pole switch rated 2.5A at 125VAC. At front-left is a four-position, single-pole switch rated 0.3A at 16VDC or 100VAC. At front right is a two-position, two-pole switch with the same rating as the one beside it. All the sealed switches allow a choice of panel mounting or through-hole printed circuit board mounting.

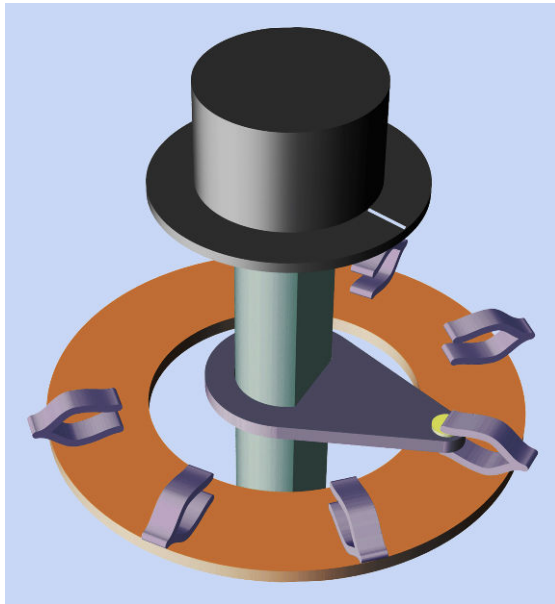


Figure 7-2. A simplified rendering of interior parts in a basic SP6T rotary switch. Arbitrary colors have been added for clarity.

How It Works

A switch may have multiple poles, each connecting with its own rotor. The rotors are likely to be on separate *decks* of the switch, but two, three, or four rotors, pointing in different directions, may be combined on a single deck if the switch has only a small number of positions.

Rotary switches are usually made with a maximum of twelve positions, but include provision for limiting the number of positions with a *stop*. This is typically a pin, which may be attached to a washer that fits around the bushing of the switch. The pin is inserted into a choice of holes to prevent the switch from turning past that point. For example, an eight-position rotary switch can be configured so that it has only seven (or as few as two) available positions.

A specification for a rotary switch usually includes the angle through which the switch turns between one position and the next. A twelve-position switch usually has a 30-degree turn angle.



Figure 7-3. A selection of rotary switches. See text for details.

Variants

Conventional

The traditional style of rotary switch is designed to be panel-mounted, with a body that ranges from 1" to 1.5" in diameter. If there is more than one deck, they are spaced from each other by about 0.5". The switch makes an audible and tactile "click" as it is turned from one position to the next.

A rugged sealed five-deck rotary switch is shown in [Figure 7-4](#). It has five poles (one per deck), and a maximum of 12 positions. The contacts are rated 0.5A at 28VDC. This type of heavy-duty component is becoming relatively rare.

If the rotor in a switch establishes a connection with the next contact a moment before breaking the connection with the previous contact, this is known as a *shorting* switch, which may also be described as a *make-before-break* switch. In a *nonshorting* or *break-before-make* switch, a tiny



Figure 7-4. A five-pole, twelve-position rotary switch.

interval separates one connection from the next. This can be of significant importance, depending on the components that are connected with the switch.

The shaft may be round, splined, or D-shaped in section. A knob is seldom supplied with a switch and must be chosen to match the shaft. Some shaft dimensions are metric, while others are measured in inches, with 1/4" diameter being the oldest standard. Some switches with a splined shaft are supplied with an adapter for a knob of D-shaped internal section; the adapter can be slipped onto the shaft in any of 12 or more positions, to minimize the inconvenience of positioning the body of the switch itself so that the knob is correctly oriented in relation to positions printed on the face of the panel.

Miniature rotary switches may be as small as 0.5" diameter, and usually terminate in pins for through-hole mounting on a PC board. Miniature switches usually have lower current ratings than full-size switches.

Rotary switches must be securely anchored to resist the high turning forces that can be inflicted

upon them by users. In a panel-mount design, a nut is tightened around a thread on the bushing of the switch. Through-hole versions can be secured to the PC board with the shaft protruding loosely through a cutout in the panel. To minimize mechanical stress on the circuit board, the detents in a PC-board switch are usually weaker than in a full-size switch, and the knob is usually smaller, allowing less leverage.

Rotary DIP

A conventional *DIP switch* is a linear array of miniature SPST switches designed to fit a standard DIP (dual-inline package) layout of holes in a circuit board. It is described in the **switch** entry of this encyclopedia. A *rotary DIP* switch (also known as an *encoded output rotary switch* or a *coded rotary switch*) does not conform with a DIP layout, despite its name. It is approximately 0.4" square and usually has five pins, one of which can be considered the input or common pin while the other four can function as outputs. The pins are spaced at 0.1" pitch from one another. Pin function and layout are not standardized.

A dial on top of the switch has either 10 positions (numbered 0 through 9) or 16 positions (0 through 9 followed by letters A through F). One switch of each type is shown in [Figure 7-5](#).

Each position of the dial closes pairs of contacts inside the component to create a unique binary-coded decimal pattern (in a 10-position switch) or binary-coded hexadecimal pattern (in the 16-position switch) on the four output pins. The pin states are shown in [Figure 7-6](#). A rotary DIP switch is a relatively flimsy device, and is not designed for frequent or heavy use. It is more likely to be a "set it and forget it" device whose state is established when it is installed in a circuit board.

Because each position of the switch is identified with a unique binary pattern, this is an example of *absolute encoding*. By contrast, a typical **rotational encoder** uses *relative encoding*, as it merely generates a series of undifferentiated pulses when the shaft is turned.

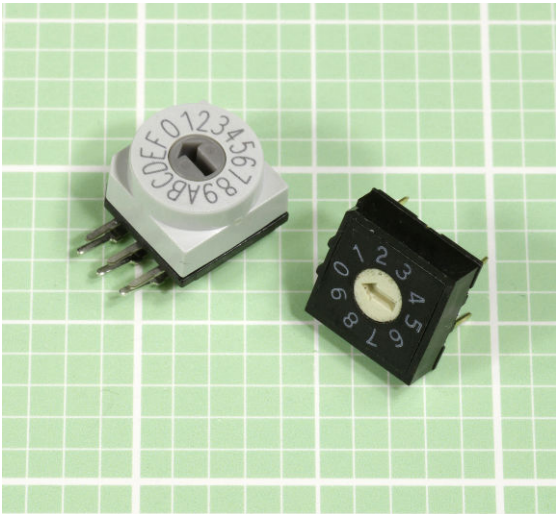


Figure 7-5. A rotary DIP switch, also known as an encoded output rotary switch, may be used as a substitute for a DIP switch in some applications.

A *real-coded* rotary DIP makes a connection between input and output pins wherever a binary 1 would exist. In the *complement-coded* version, the output is inverted. The switch is primarily intended for use with a **microcontroller**, enabling only four binary input pins on the microcontroller to sense up to sixteen different switch positions.

A six-pin rotary DIP variant is available from some manufacturers, with two rows of three pins, the two center pins in each row being tied together internally, and serving as the pole of the switch.

Rotary DIPs are available with a screw slot, small knurled knob, or larger knob. The screw-slot version minimizes the height of the component, which can be relevant where circuit boards will be stacked close together. A *right-angle PC* variant stands at 90 degrees to the circuit board, with pins occupying a narrower footprint. The switch on the left in [Figure 7-5](#) is of this type.

While most rotary DIPs are through-hole components, surface-mount versions are available.

Most rotary DIPs are sealed to protect their internal components during wave-soldering of circuit boards.

Switch Position	Pin 1	Pin 2	Pin 3	Pin 4
0	●	●	●	●
1	●	●	●	●
2	●	●	●	●
3	●	●	●	●
4	●	●	●	●
5	●	●	●	●
6	●	●	●	●
7	●	●	●	●
8	●	●	●	●
9	●	●	●	●
A	●	●	●	●
B	●	●	●	●
C	●	●	●	●
D	●	●	●	●
E	●	●	●	●
F	●	●	●	●

Figure 7-6. Positive and negative states of the four output pins of a *real-coded* 16-position rotary DIP switch, assuming that the common pin of the switch is connected with a positive supply voltage. A ten-position rotary DIP switch would use only the states from 0 through 9. In a *complement-coded* switch, the positive and negative states would be reversed.

Gray Code

A Gray code (named after its originator, Frank Gray) is a system of absolute encoding of a switch output, using a series of nonsequential binary numbers that are chosen in such a way that each number differs by only one digit from the preceding number. Such a series is useful because it eliminates the risk that when a switch turns, some bits in the output will change before others, creating the risk of erroneous interpretation. A minority of rotary switches or rotational en-

coders are available with Gray-coded outputs. Typically, a microcontroller must use a lookup table to convert each binary output to an angular switch position.

PC Board Rotary Switch

Miniature switches with a conventional, non-encoded output are available for printed-circuit board mounting, sometimes requiring a screwdriver or hex wrench to select a position. A single-pole eight-position switch of this type is shown in [Figure 7-7](#). Its contacts are rated to carry 0.5A at 30VDC, but it is not designed to switch this current actively.

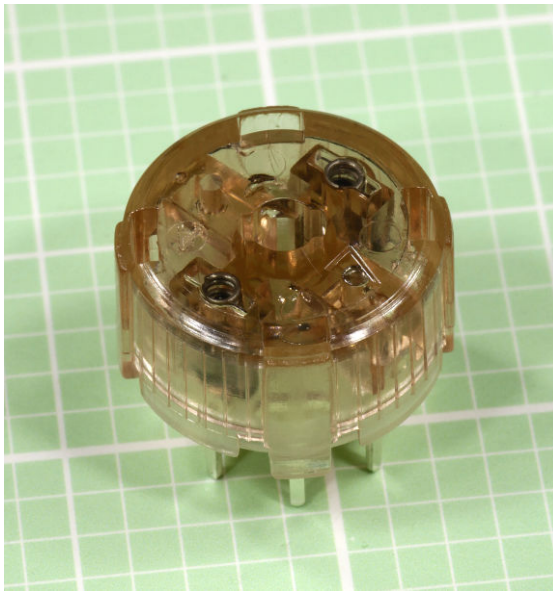


Figure 7-7. This miniature switch is designed for insertion on a printed circuit board. It can be used to make a setting before a device is shipped to the end user.

Mechanical Encoder

A [mechanical encoder](#) functions similarly to a rotary DIP switch but is intended for much heavier use. It outputs a binary-coded-decimal value corresponding with its shaft position, is typically the size of a miniature rotary switch, and is designed

for panel mounting. The Grayhill Series 51 allows 12 positions, each generating a code among four terminals. The Bourns EAW provides 128 positions, each generating a code among 8 terminals.

Pushwheel and Thumbwheel

A [pushwheel switch](#) is a simple electromechanical device that enables an operator to provide a code number as input to data processing equipment, often in industrial process control. The decimal version contains a wheel on which numbers are printed, usually in white on black, from 0 through 9, visible one at a time through a window in the face of the switch. A button above the wheel, marked with a minus sign, rotates it to the next lower number, while a button below the wheel, marked with a plus sign, rotates it to the next higher number. A connector at the rear of the unit includes a common (input) pin and four output pins with values 1, 2, 4, and 8. An additional set of pins with values 1, 2, 3, and 4 is often provided. The states of the output pins sum to the value that is currently being displayed by the wheel. Often two, three, or four pushwheels (each with an independent set of connector pins) are combined in one unit, although individual pushwheels are available and can be stacked in a row.

A [thumbwheel switch](#) operates like a pushwheel switch, except that it uses a thumbwheel instead of two buttons. Miniaturized thumbwheel switches are available for through-hole mounting on PC boards.

Hexadecimal versions are also available, with numbers from 0 through 9 followed by letters A through F, although they are less common than decimal versions.

Keylock

A [keylock switch](#) is generally a two-position rotary switch that can be turned only after insertion of a key in a lock attached to the top of the shaft. This type of switch almost always has an OFF-(ON) configuration and is used to control power.

Keylock switches are found in locations such as elevators, for fire-department access; in cash registers; or on data-processing equipment where switching power on or off is reserved for a system administrator.

Values

A full-size rotary switch may be rated from 0.5A at 30VDC to 5A at 125VAC, depending on its purpose. A very few switches are rated 30A at 125VAC; these are high-quality, durable, expensive items.

A typical rotary DIP switch is rated 30mA at 30VDC and has a carrying current rating (continuous current when no switching occurs) of no more than 100mA at 50VDC.

How to Use it

In addition to its traditional purpose as a mode or option selector, a rotary switch provides a user-friendly way to input data values. Three ten-position switches, for instance, can allow user input of a decimal number ranging from 000 to 999.

When used with a **microcontroller**, a rotary switch can have a *resistor ladder* mounted around its contacts, like a multi-point *voltage divider*, so that each position of the rotor provides a unique potential ranging between the positive supply voltage and negative ground. This concept is illustrated in [Figure 7-8](#), where all the resistors have the same value. The voltage can be used as an input to the microcontroller, so long as the microcontroller shares a common ground with the switch. An analog-digital converter inside the microcontroller translates the voltage into a digital value. The advantage of this scheme is that it allows very rapid control by the user, while requiring only one pin on the microcontroller to sense as many as twelve input states.

For a ladder consisting of 8 resistors, as shown, each resistor could have a value of 250Ω. (The specifications for a particular microcontroller

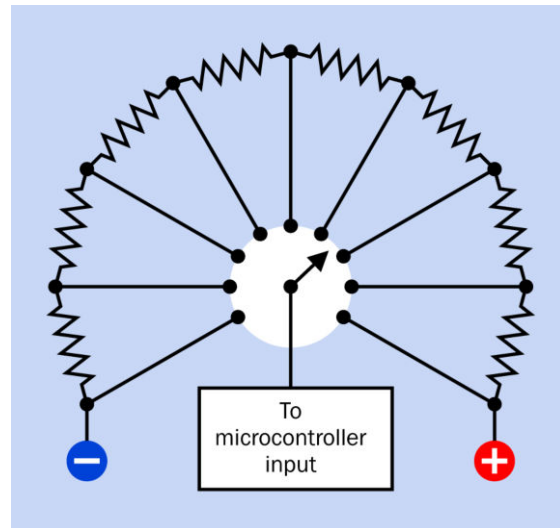


Figure 7-8. A *resistor ladder* can be formed around the contacts of a rotary switch, with the pole of the switch connected to a microcontroller that has an analog-digital converter built in. The microcontroller converts the voltage input to an internal digital value. Thus, one pin can sense as many as twelve input states.

might require other values.) To avoid ambiguous inputs, a nonshorting rotary switch should be used in this scheme. A *pullup resistor* of perhaps 10K should be added to the microcontroller input, so that there is no risk of it “floating” when the switch rotor is moving from one contact to the next. The code that controls the microcontroller can also include a blanking interval during which the microcontroller is instructed to ignore the switch.

Because the rotary switch is an electromechanical device, it has typical vulnerabilities to dirt and moisture, in addition to being bulkier, heavier, and more expensive than a **rotational encoder**. Rotary switches have also been partially replaced by pushbuttons wired to a microcontroller. This option is found on devices ranging from digital alarm clocks to cellular phones. In addition to being cheaper, the pushbutton alternative is preferable where space on a control panel, and behind it, is limited.

What Can Go Wrong

Vulnerable Contacts

Most modern rotary switches are sealed, but some are not. Any switch with exposed contacts will be especially vulnerable to dirt and moisture, leading to unreliable connections. This was an issue in old-fashioned TV sets, where periodic contact cleaning of the channel selector switch was needed.

Exposed contacts are also more vulnerable to side-effects from temperature cycling (when a device warms up and then cools down).

Contact Overload

The contacts on a cheap rotary switch are especially vulnerable to arcing, as the user may turn the switch slowly, causing gradual engagement and disengagement of contacts instead of the snap-action that is characteristic of a well-made toggle switch. If a rotary switch may control significant currents or current surges, it must be appropriately rated, regardless of the extra expense. For more information on arc suppression in switches, see [“Arcing”](#) (page 47).

Misalignment

Most knobs for rotary switches consist of a pointer, or have a white line engraved to provide clear visual indication of the position of the switch. If this does not align precisely with indications printed on the panel, confusion will result. For hand-built equipment, the switch can be installed first, after which the control-panel indications can be glued or riveted in place on a separate piece of laminated card, plastic, or metal for precise alignment. If the switch is not secured tightly, its body may turn slightly under repeated stress, leading to erroneous interpretation of the knob position.

Misidentified Shorting Switch

If a shorting switch is used where a nonshorting switch was intended, the results can be disconcerting or even destructive, as one terminal will be briefly connected with the adjacent terminal while the switch is being turned. Multiple functions of a circuit may be activated simultaneously, and in a worst case scenario, adjacent terminals may be connected to opposite sides of the same power supply.

User Abuse

The turning force that must be applied to a full-size conventional rotary switch is significantly greater than the force that is applied to most other types of panel-mounted switches. This encourages aggressive treatment, and the turning motion is especially likely to loosen a nut holding the switch in place. The lighter action characteristic of miniature rotary switches does not necessarily solve this problem, as users who are accustomed to older-style switches may still apply the same force anyway.

Rotary switches should be mounted in expectation of rough use. It is prudent to use Loc-Tite or a similar compound to prevent nuts from loosening, and a switch should not be mounted in a thin or flimsy panel. When using a miniature rotary switch that has through-hole mounting in a circuit board, the board must be sufficiently robust and properly secured.

Wrong Shaft, Wrong Knobs, Nuts That Get Lost, Too Big to Fit

These problems are identical to those that can be encountered with a **potentiometer**, which are discussed in that entry in this encyclopedia.



rotational encoder

The term *rotational encoder* used to be reserved for high-quality components, often using optical methods to measure rotation with precision (more than 100 intervals in 360 degrees). Cheaper, simpler, electromechanical devices were properly referred to as *control shaft encoders*. However, the term *rotational encoder* is now applied to almost any device capable of converting rotational position to a digital output via opening and closing internal mechanical contacts; this is the sense in which the term is used here. It is sometimes distinguished from other types of encoder with the term *mechanical rotary encoder*. Magnetic and optical rotary encoders do not contain mechanical switches, are classified as *sensors* by this encyclopedia, and will appear in Volume 3. They are found in a device such as an *optical mouse*.

OTHER RELATED COMPONENTS

- **rotary switch** (See [Chapter 7](#))

What It Does

A rotational encoder has a knob that a user can turn to display a series of prompts on an LCD screen, or to adjust the input or output on a product such as a stereo receiver. The component is almost always connected to inputs on a **microcontroller** and is usually fitted with *de-tents* that provide tactile feedback suggesting many closely spaced positions. The encoder often allows the user to make a selection by pushing the knob in, which closes an internal momentary switch. Thus, this type of encoder functions as a pushbutton as well as a switch.

A rotational encoder is an *incremental* or *relative* device, meaning that it merely creates and breaks internal switch connections when rotation occurs, without providing a unique code to identify each absolute rotational position. An *absolute* encoder is discussed in the **rotary switch** entry of this encyclopedia.

No schematic symbol exists to represent a rotational encoder.

How It Works

An encoder contains two pairs of contacts, which open and close out of phase with each other when the shaft rotates. In a clockwise direction, the A pair of contacts may be activated momentarily before the B pair; in a counter-clockwise direction, the B pair may be activated before the A pair. (Some encoders reverse this phase difference.) Thus if one contact from each pair is connected with two inputs of an appropriately programmed microcontroller, and if the other contact of each pair is connected with negative ground, the microcontroller can deduce which way the knob is turning by sensing which pair of contacts closes first. The microcontroller can then count the number of pulses from the contacts and interpret this to adjust an output or update a display.

A simplified schematic is shown in [Figure 8-1](#). The two buttons inside the dashed line represent the two pairs of contacts inside the encoder, while the chip is a microcontroller. The knob and shaft that activate the internal switches are not shown. The schematic assumes that when a contact closes, it pulls the chip input to a low state. A pullup resistor is added to each input of the chip to prevent the pins from “floating” when either pair of contacts is open.

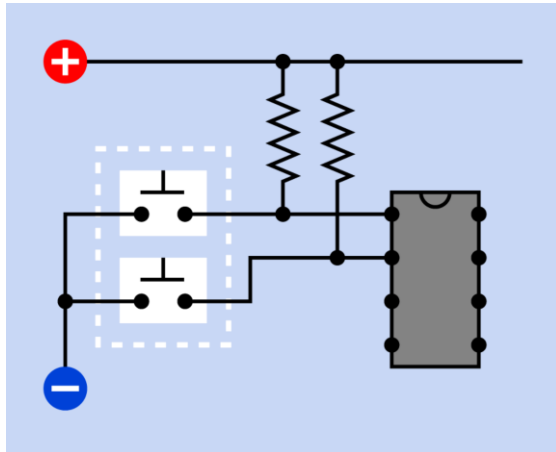


Figure 8-1. Simplified schematic showing the typical set-up for a rotational encoder. The pushbuttons inside the dashed line represent the contacts inside the encoder. The chip is a microcontroller.

[Figure 8-2](#) gives a conceptual view of the outputs of an encoder that is turned clockwise (top) and then counter-clockwise (bottom). Some encoders may reverse this phase sequence. Red and black colors have been assigned to the pin states on the assumption that the terminals that are common to both pairs of contacts are connected with negative ground. Thus a “high” pulse in the graphical representation actually indicates that the encoder is grounding its output.

Microcontrollers have become so ubiquitous, and rotational encoders are so cheap, they have displaced rotary switches in many applications where a low current is being switched. The com-

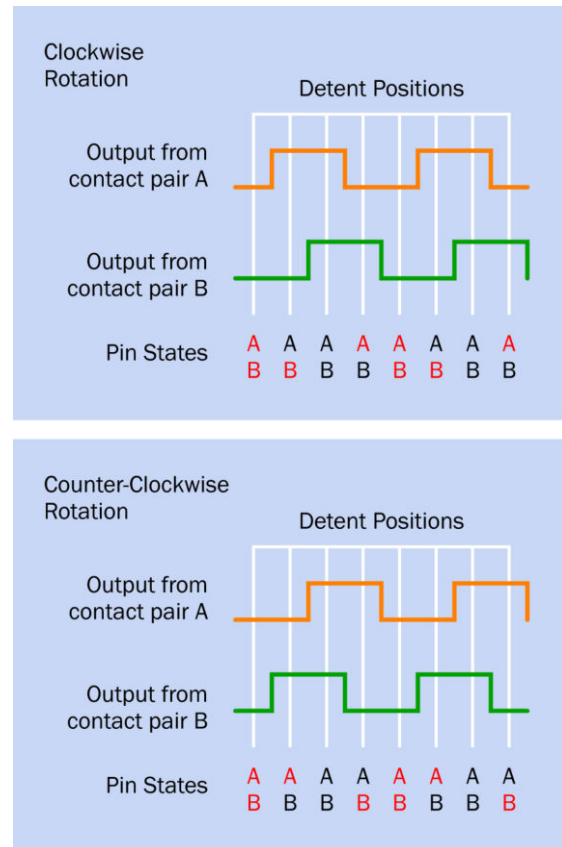


Figure 8-2. Hypothetical outputs from a rotational encoder, assuming that the common terminals of the contact pairs are connected to negative ground. A high pulse in the graphical representation therefore indicates that the contact pair is grounded. The number of detents relative to the number of pulses per rotation varies from one type of encoder to another.

bination of a rotational encoder and a microcontroller is very versatile, allowing display and control of an almost unlimited number of menus and options.

Variants

There are two types of rotational encoders containing mechanical contacts: absolute and relative. An absolute encoder generates a code corresponding with each specific rotational position. The code is usually a binary output among

four or more pins. It is discussed under [mechanical encoder](#) in the **rotary switch** section of this encyclopedia. The variants listed here are all relative encoders.

Pulses and Detents

Rotational encoders from different manufacturers may have as few as 4 or as many as 24 pulses per rotation (PPR), with 12 to 36 detents (or no detents at all, in a few models.) The relationship between pulses and detents shown in [Figure 8-2](#) is typical but is far from being universal. The number of detents may be equal to, greater than, or less than the number of pulses per rotation.

Format

Rotational encoders are generally panel-mounted or through-hole devices. In the latter category, most are horizontally mounted, with a minority being at 90 degrees to the board.

Output

In an encoder containing two switches, four switch-state combinations are possible: OFF-OFF, ON-OFF, OFF-ON, and ON-ON. This is known as a [quadrature](#) output. All of the rotational encoders discussed here conform with that system.

Rotational Resistance

Rotational encoders vary widely in the resistance that they offer when the user turns the knob. This is largely a function of the detents, if they are included. Still, all rotational encoders generally offer less rotational resistance than a rotary switch, and do not have the kind of heavy-duty knobs that are typically used with rotary switches. Since an encoder creates only a stream of pulses without any absolute positional information, a knob with any kind of pointer on it is inappropriate.

Values

Virtually all rotational encoders are designed to work with a low-voltage supply, 12VDC or less. All of them are intended for low currents, reflect-

ing their purpose to drive microcontroller inputs. Some sample rotational encoders are pictured in [Figure 8-3](#). At rear: nine pulses per rotation (PPR), 36 detents, 10mA at 10VDC. Far left: 20PPR, 20 detents, with switch. Far right: 24PPR, no detents, 1mA at 5VDC. Center (blue): 16PPR, no detents, 1mA at 5VDC. Front: 12PPR, 24 detents, 1mA at 10VDC, requires Allen wrench or similar hexagonal shaft to engage with the rotor.



Figure 8-3. Rotational encoders with a variety of specifications. See text for details.

Contact Bounce

Any mechanical switch will suffer some degree of [contact bounce](#) when its contacts close. Data-sheets for rotational encoders may include a specification for bounce duration ranging from around 2ms to 5ms, which is sometimes known as the [settling time](#). Naturally, a lower value is preferred. The microcontroller that interprets the positional information from the encoder can include a debouncing routine that simply disregards any signals during the bounce period following switch closure.

Sliding Noise

Sliding noise is the opposite of contact bounce. When two contacts have made a connection and then rub across each other (as occurs inside a rotational encoder while the knob is being turned), the connection may suffer momentary lapses. Datasheets for rotational encoders generally do not supply ratings for this.

How to Use it

As noted above, a rotational encoder can only be used in conjunction with a microcontroller or similar device that is capable of interpreting the phase difference between the pairs of contacts, and is capable of counting the number of opening/closing events while the knob is being turned. (Some dedicated chips are designed for this specific purpose.)

It can be adapted to be driven by a **stepper motor**, to provide feedback regarding the rotation of the motor shaft, and its output can also be interpreted to calculate angular acceleration.

Programming the microcontroller is the most significant obstacle. Generally the program should follow a sequence suggested by this pseudocode:

Check:

- If the encoder contains a pushbutton switch, check it. If the pushbutton is being pressed, go to an appropriate subroutine.
- The status of contacts A.
- The status of contacts B.

Compare their status with previously saved states for A and B. If the status has not changed, repeat from Check.

Debounce:

- Recheck the contacts status rapidly and repeatedly for 50ms, and count the states for

contacts A and B. (The 50ms duration may be adjusted for different encoders, as an encoder with a higher number of pulses per rotation will tend to create shorter pulses.)

- Compare the total number of changed states with unchanged states.

If the changed states are in a small minority, probably the signal was erroneous, caused by bounce or sliding noise. Go back to Check and start over.

Interpret:

- Deduce the rotational direction from these four possibilities:
 - Contacts A were open and have closed.
 - Contacts A were closed and have opened.
 - Contacts B were open and have closed.
 - Contacts B were closed and have opened. (The specific type of encoder will determine how these transitions are interpreted.)
- Revise the variable storing the direction of rotation if necessary.
- Depending on the direction of rotation, increment or decrement a variable that counts pulses.
- Take action that is appropriate to the direction of rotation and the cumulative number of pulses.
- Go back to Check again.

What Can Go Wrong

Switch Bounce

In addition to a debouncing algorithm in the microcontroller, a 0.1μF bypass capacitor can be used with each of the output terminals from the encoder, to help reduce the problem of switch bounce.

Contact Burnout

Rotational encoders are TTL-compatible. They are not generally designed to drive even a small output device, such as an LED. The contacts are extremely delicate and will be easily damaged by any attempt to switch a significant current.

relay



Properly known as an *electromagnetic armature relay* to distinguish it from a *solid-state relay*. However, the full term is very rarely used. It may also be described as an *electro-mechanical relay*, but the term **relay** is normally understood to mean a device that is not solid state.

OTHER RELATED COMPONENTS

- **solid state relay** (Volume 2)
- **switch** (See [Chapter 6](#))

What It Does

A relay enables a signal or pulse of electricity to switch on (or switch off) a separate flow of electricity. Often, a relay uses a low voltage or low current to control a higher voltage and/or higher current. The low voltage/low current signal can be initiated by a relatively small, economical switch, and can be carried to the relay by relatively cheap, small-gauge wire, at which point the relay controls a larger current near to the load. In a car, for example, turning the ignition switch sends a signal to a relay positioned close to the starter motor.

While solid-state switching devices are faster and more reliable, relays retain some advantages. They can handle double-throw and/or multiple-pole switching and can be cheaper when high voltages or currents are involved. A comparison of their advantages relative to **solid state relays** and **transistors** is tabulated in the entry on **bipolar transistor** in [Figure 28-15](#).

Common schematic symbols for single-throw relays are shown in [Figure 9-1](#) and for double-

throw relays in [Figure 9-2](#). The appearance and orientation of the coil and contacts in the symbols may vary significantly, but the functionality remains the same.

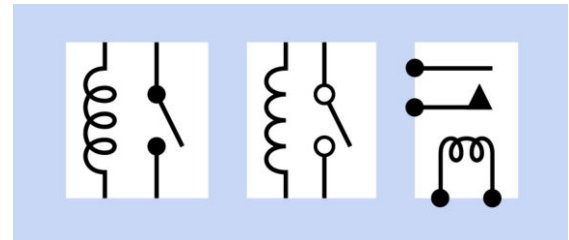


Figure 9-1. Commonly used schematic symbols for a SPST relay. The symbols are functionally identical.

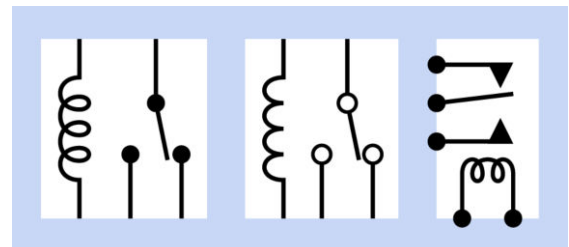


Figure 9-2. Commonly used schematic symbols for a SPDT relay. The symbols are functionally identical.

How It Works

A relay contains a *coil*, an *armature*, and at least one pair of *contacts*. Current flows through the coil, which functions as an **electromagnet** and generates a magnetic field. This pulls the armature, which is often shaped as a pivoting bracket that closes (or opens) the contacts. These parts are visible in the simplified rendering of a DPST relay in [Figure 9-3](#). For purposes of identification, the armature is colored green, while the coil is red and the contacts are orange. The two blue blocks are made of an insulating material, the one on the left supporting the contact strips, the one on the right pressing the contacts together when the armature pivots in response to a magnetic field from the coil. Electrical connections to the contacts and the coil have been omitted for simplicity.

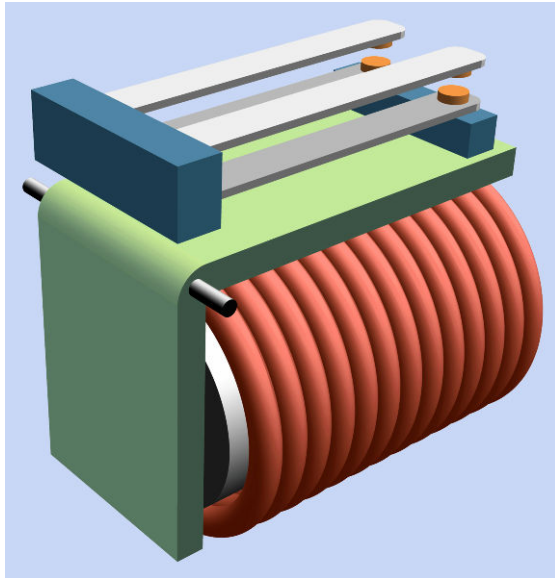


Figure 9-3. This simplified rendering shows the primary parts of a DPST relay. See text for details.

Various small relays, capable of handling a variety of voltages and currents, are pictured in [Figure 9-4](#). At top-left is a 12VDC automotive relay, which plugs into a suitable socket shown immediately below it. At top-right is a 24VDC SPDT

relay with exposed coil and contacts, making it suitable only for use in a very clean, dry environment. Continuing downward, the four sealed relays in colored plastic cases are designed to switch currents of 5A at 250VAC, 10A at 120VAC, 0.6A at 125VAC, and 2A at 30VDC, respectively. The two blue relays have 12VDC coils, while the red and yellow relays have 5V coils. All are nonlatching, except for the yellow relay, which is a latching type with two coils. At bottom-left is a 12VDC relay in a transparent case, rated to switch up to 5A at 240VAC or 30VDC.

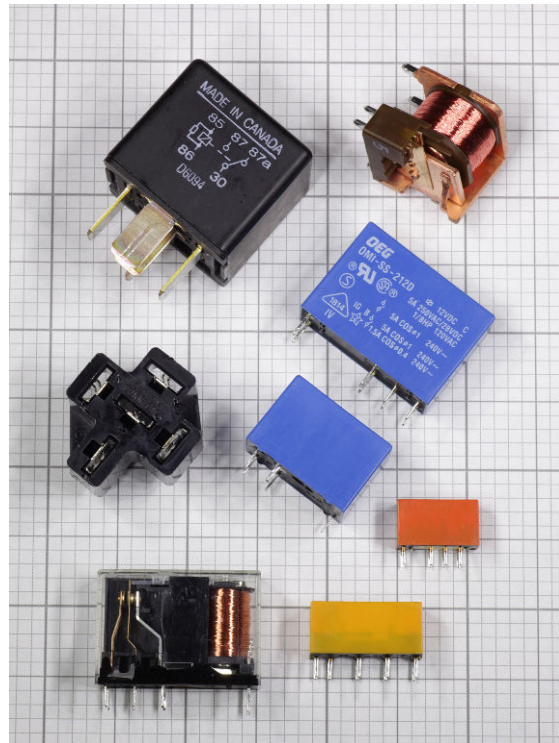


Figure 9-4. An assortment of small DC-powered relays. See text for details.

The configuration of a relay is specified using the same abbreviations that apply to a **switch**. SP, DP, 3P, and 4P indicate 1, 2, 3, or 4 poles (relays with more than 4 poles are rare). ST and DT indicate single-throw or double-throw switching. These abbreviations are usually concatenated, as in 3PST or SPDT. In addition, the terminology Form

A (meaning normally open), Form B (normally closed), and Form C (double-throw) may be used, preceded by a number that indicates the number of poles. Thus “2 Form C” means a DPDT relay.

Variants

Latching

There are two basic types of relay: *latching* and *nonlatching*. A nonlatching relay, also known as a *single side stable* type, is the most common, and resembles a *momentary switch* or *pushbutton* in that its contacts spring back to their default state when power to the relay is interrupted. This can be important in an application where the relay should return to a known state if power is lost. By contrast, a latching relay has no default state. Latching relays almost always have double-throw contacts, which remain in either position without drawing power. The relay only requires a short pulse to change its status. In semiconductor terms, its behavior is similar to that of a *flip-flop*.

In a *single-coil* latching relay, the polarity of voltage applied to the coil determines which pair of contacts will close. In a *dual-coil* latching relay, a second coil moves the armature between each of its two states.

Schematic symbols for a dual-coil latching relay are shown in Figure 9-5. Some symbol styles do not make it clear which switch position each coil induces. It may be necessary to read the manufacturer’s datasheet or test the relay by applying its rated voltage to randomly selected terminal pairs while testing for continuity between other terminal pairs.

Polarity

There are three types of DC relay. In a *neutral relay*, polarity of DC current through the coil is irrelevant. The relay functions equally well either way. A *polarized relay* contains a diode in series with the coil to block current in one direction. A *biased relay* contains a permanent magnet near the armature, which boosts performance when

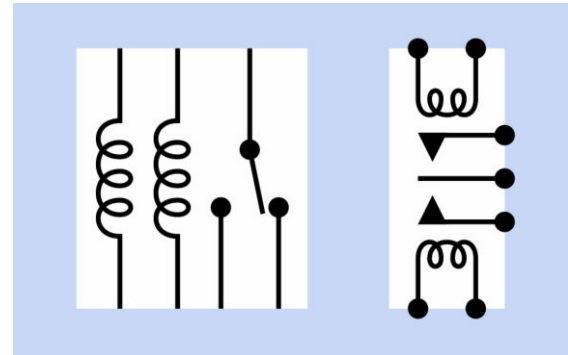


Figure 9-5. Schematic symbols for a two-coil latching relay. The symbols are functionally identical.

current flows through the coil in one direction, but blocks a response when the current flows through the coil in the opposite direction. Manufacturers’ datasheets may not use this terminology, but will state whether the relay coil is sensitive to the polarity of a DC voltage.

All relays can switch AC current, but only an *AC relay* is designed to use AC as its coil current.

Pinout Variations

The layout and function of relay pins or quick connects is not standardized among manufacturers. Often the component will have some indication of pin functions printed on it, but should always be checked against the manufacturer’s datasheet and/or tested for continuity with a meter.

Figure 9-6 shows four sample pin configurations, adapted from a manufacturer’s datasheet. These configurations are functionally quite different, although all of them happen to be for DPDT relays. In each schematic, the coil of the relay is shown as a rectangle, while the pins are circles, black indicating an energized state and white indicating a non-energized state. The bent lines show the possible connections between the poles and other contacts inside the relay. The contacts are shown as arrows. Thus, pole 4 can connect with either contact 3 or contact 5, while pole 9 can connect with either contact 8 or contact 10.

Top-left: Polarized nonlatching relay in its resting condition, with no power applied. Top right: Single-coil latching relay showing energized contacts (black circles) when the coil is powered with the polarity indicated. If the polarity is reversed, the relay flips to its opposite state. Some manufacturers indicate the option to reverse polarity by placing a minus sign alongside a plus sign, and a plus sign alongside a minus sign. Bottom-left and bottom-right: Polarized latching relays with two coils, with different pinouts.

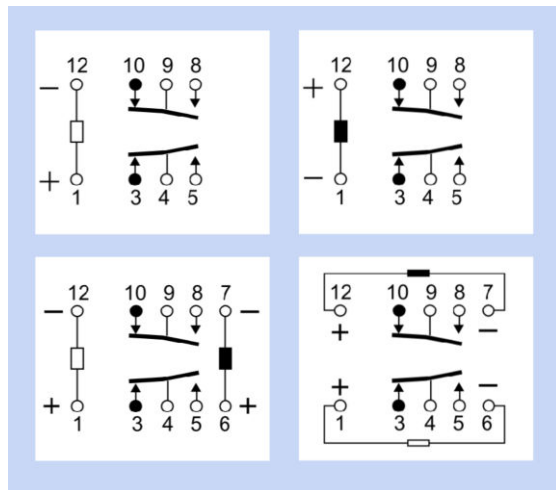


Figure 9-6. Relay pinouts depicted in the style commonly found in manufacturers' datasheets, showing different relay types. Top-left: Single coil, nonlatching. Top-right: Single coil, latching. Bottom left: Two-coil, latching. Bottom right: Two-coil, latching, alternate pinouts. (Adapted from a Panasonic datasheet.)

In these diagrams, the relay is seen from above. Some datasheets show the relay seen from below, and some show both views. Some manufacturers use slightly different symbols to indicate interior functions and features. When in doubt, use a meter for verification.

Reed Relay

A **reed relay** is the smallest type of electromechanical relay with applications primarily in test equipment and telecommunications. With a coil resistance ranging from 500 to 2000 ohms, these relays consume very little power. The design con-

sists of a **reed switch** with a coil wrapped around it. **Figure 9-7** shows a simplified rendering. The two black contacts are enclosed in a glass or plastic envelope and magnetized in such a way that a magnetic field from the surrounding coil bends them together, creating a connection. When power to the coil is disconnected, the magnetic field collapses and the contacts spring apart.

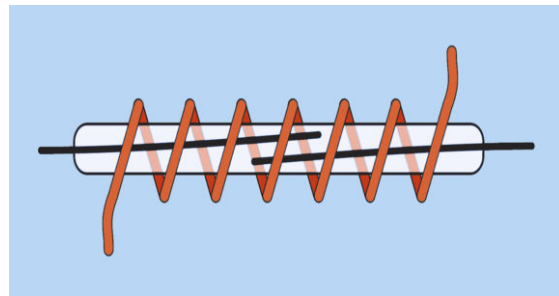


Figure 9-7. This simplified rendering shows a reed relay, consisting of a magnetized reed switch inside a glass or plastic pod, activated by a coil wrapped around it.

In **Figure 9-8**, two reed relays are shown, at top-left and center-right. At bottom-left, the type of relay on the right has been opened by a belt sander to reveal its copper coil and inside that, a capsule in which the relay contacts are visible.

Surface-mount reed relays can be smaller than 0.5" × 0.2". Through-hole versions are often around 0.7" × 0.3" with pins in two rows, though some are available in SIP packages.

Reed relays have limited current switching capacity and are not suitable to switch inductive loads.

Small Signal Relay

A **small signal relay** is also known as a **low signal relay**. This type may have a footprint as small as a reed relay but generally stands slightly taller, requires slightly more coil current, and is available in versions that can switch slightly higher



Figure 9-8. Three reed relays, one of which has had its packaging partially removed by a belt sander to reveal its copper coil and internal contacts.

voltages and currents. There are usually two rows of pins, spaced either 0.2" or 0.3" apart. The red and orange relays in [Figure 9-4](#) are small signal relays.

Automotive Relays

An *automotive relay* is typically packaged in a cube-shaped black plastic case with quick-connect terminals at the bottom, typically plugged into a socket. Naturally they are designed to switch, and be switched by, a 12VDC supply.

General Purpose/Industrial

These relays cover a very wide range and are usually built without significant concern for size. They may be capable of switching high currents at high voltages. Typically they are designed to plug into a socket such as an *octal base* of the type that was once used for **vacuum tubes**. The base, in turn, terminates in solder tabs, screws, or quick connects and is designed to be screwed to a chassis. It allows the relay to be unplugged and swapped without resoldering.

Two industrial relays are shown in [Figure 9-9](#). Both are DPDT type with 12VDC coils and rated to switch up to 10A at 240VAC. The one on the left has an octal base. An octal socket that fits an octal base is shown in [Figure 9-10](#).

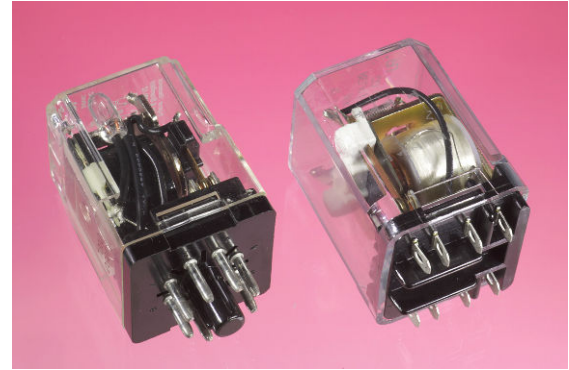


Figure 9-9. Two relays powered by 12VDC, capable of switching up to 10A at 240VAC.



Figure 9-10. An octal socket with screw terminals, designed to accept a relay with an octal base.

Time Delay Relay

Generally used to control industrial processes, a *time delay relay* switches an output on and off at

preset time intervals that can be programmed to repeat. The example in [Figure 9-11](#) has a 12VDC coil and is rated to switch up at 10A at 240VAC. It has an octal base.



Figure 9-11. The control switches on a time-delay relay, allowing separately configured “on” and “off” intervals.

Contactor

A contactor functions just like a relay but is designed to switch higher currents (up to thousands of amperes) at higher voltages (up to many kilovolts). It may range from being palm-sized to measuring more than one foot in diameter, and may be used to control heavy loads such as very large motors, banks of high-wattage lights, and heavy-duty power supplies.

Values

Datasheets usually specify maximum voltage and current for the contacts, and nominal voltage and current for the coil, although in some cases the coil resistance is stated instead of nominal coil current. The approximate current consumption can be estimated, if necessary, by us-

ing Ohm’s Law. The minimum voltage that the relay needs for activation is sometimes described as the *Must Operate By* voltage, while the *Must Release By* voltage is the maximum coil voltage that the relay will ignore. Relays are rated on the assumption that the coil may remain energized for long periods, unless otherwise stated.

While the contact rating may suggest that a relay can switch a large load, this is not necessarily true if the load has significant inductance.

Reed relays

Usually use a coil voltage of 5VDC and have a contact rating of up to 0.25A at 100V. Through-hole (PCB) versions may have a coil voltage of 5VDC, 6VDC, 12VDC, or 24VDC and in some cases claim to switch 0.5A to 1A at up to 100V, although this rating is strictly for a noninductive load.

Small signal/low signal relays

Usually use a coil voltage ranging from 5VDC to 24VDC, drawing about 20mA. Maximum switching current for noninductive loads ranges from 1A to 3A.

Industrial/general purpose relays

A very wide range of possible values, with coil voltages ranging up to 48VDC or 125VAC to 250VAC. Contact rating is typically 5A to 30A.

Automotive relays

Coil voltage of 12VDC, and contact rating often 5A at up to 24VDC.

Timer relays

Usually these specify a coil voltage of 12VDC, 24VDC, 24VAC, 125VAC, or 230VAC. The timed interval can range from 0.1 sec to 9999 hours in some cases. Common values for contact ratings are 5A up to 20A, with a voltage of 125V to 250V, AC or DC.

How to Use it

Relays are found in home appliances such as dishwashers, washing machines, refrigerators, air conditioners, photocopy machines, and other products where a substantial load (such as a motor or compressor) has to be switched on and off by a control switch, a thermostat, or an electronic circuit.

Figure 9-12 shows a common small-scale application in which a signal from a microcontroller (a few mA at 5VDC) is applied to the base of a transistor, which controls the relay. In this way, a logic output can switch 10A at 125VAC. Note the rectifier diode wired in parallel with the relay coil.

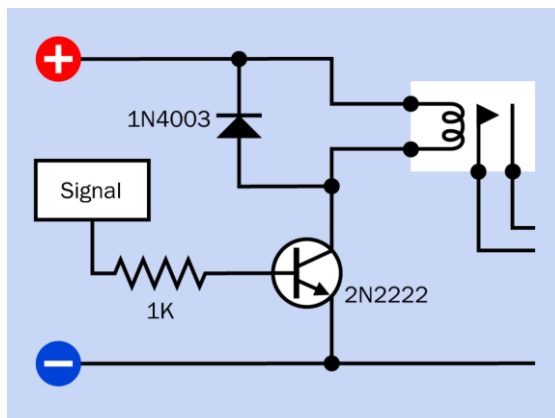


Figure 9-12. A signal from a digital source such as a microcontroller can switch substantial voltage and current if it is applied to the base of a transistor that activates a relay.

A latching relay is useful wherever a connection should persist when power is switched off or interrupted, or if power consumption must be minimized. Security devices are one common application. However, the circuit may require a “power reset” function to restore known default settings of latched relays.

A circuit including every possible protection against voltage spikes is shown in Figure 9-13, including a snubber to protect the relay contacts, a rectifier diode to suppress back-EMF generated by the relay coil, and another rectifier diode to

protect the relay from EMF generated by a motor when the relay switches it on and off. The snubber can be omitted if the motor draws a relatively low current (below 5A) or if the relay is switching a noninductive load. The diode around the relay coil can be omitted if there are no semiconductors or other components in the circuit that are vulnerable to voltage spikes. However, a spike can affect components in adjacent circuits that appear to be electrically isolated. A severe spike can even be transmitted back into 125VAC house wiring. For information on using a resistor-capacitor combination to form a snubber, see “Snubber” (page 108).

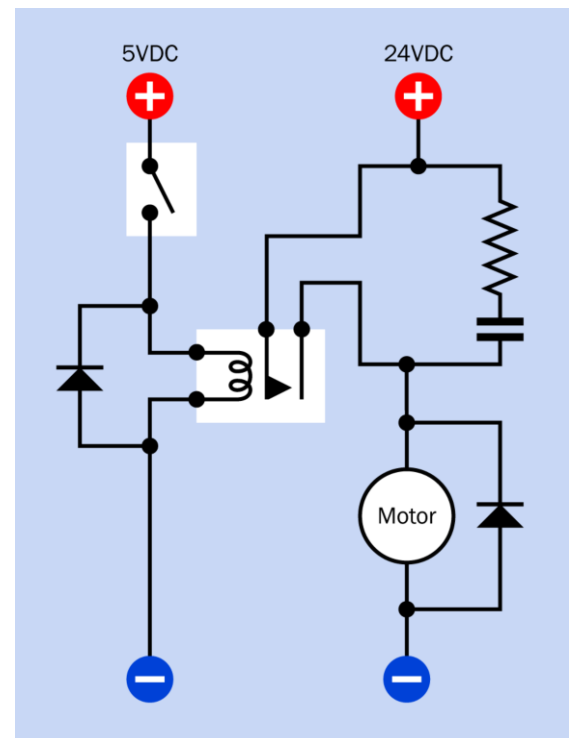


Figure 9-13. This hypothetical schematic shows three types of protection against voltage spikes induced by an inductive load (a motor, in this instance) and the coil of the relay.

What Can Go Wrong

Wrong Pinouts

The lack of standardization of relay pinouts can cause errors if one relay is replaced with another that appears to be the same, but isn't. In particular, the pins that connect with normally-closed contacts may trade places with pins connected with normally-open contacts, in relays from different manufacturers.

Pinouts are also confusing in that some data-sheets depict them from above, some from below, and some from both perspectives.

Wrong Orientation

Small relays of through-hole type usually have pins spaced in multiples of 0.1". This allows them to be inserted the wrong way around in a perforated board. Almost all relays have an identification mark molded into one end or one corner of the plastic shell. Manufacturers do not standardize the position and meaning of these marks, but they are usually replicated in datasheets. When using a relay of a type that you have not used before, it is a sensible precaution to test it with a meter to verify the functions of its terminals before installing it.

Wrong Type

A latching relay may have exactly the same appearance as a nonlatching relay from the same manufacturer, and the same two pins may energize the coil. However, in a latching relay, the contacts won't spring back to their non-energized position, causing functional errors that may be difficult to diagnose. The part numbers printed on latching and nonlatching versions of the same relay may differ by only one letter or numeral and should be checked carefully.

Wrong Polarity

A relay with a DC-energized coil may require power to be applied with correct polarity and may malfunction otherwise.

AC and DC

A relay coil designed to be powered by DC will not work from AC and vice-versa. The contact rating of a relay is likely to be different depending whether it is switching AC or DC.

Chatter

This is the noise created by relay contacts when they make rapid intermittent connection. Chatter is potentially damaging to relay contacts and should be avoided. It can also create electrical noise that interferes with other components. Likely correctible causes of chatter include insufficient voltage or power fluctuations.

Relay Coil Voltage Spike

A relay coil is an inductive device. Merely switching a large relay on and off can create voltage spikes. To address this problem, a rectifier diode should be placed across the coil terminals with polarity opposing the energizing voltage.

Arcing

This problem is discussed in the **switch** entry of this encyclopedia. See "[Arcing](#)" ([page 47](#)). Note that because the contacts inside a [reed relay](#) are so tiny, they are especially susceptible to arcing and may actually melt and weld themselves together if they are used to control excessive current or an inductive load.

Magnetic Fields

Relays generate magnetic fields during operation and should not be placed near components that are susceptible.

The reed switch inside a reed relay can be unexpectedly activated by an external magnetic field. This type of relay may be enclosed in a metal shell to provide some protection. The adequacy of this protection should be verified by testing the relay under real-world conditions.

Environmental Hazards

Dirt, oxidation, or moisture on relay contacts is a significant problem. Most relays are sealed and should remain sealed.

Relays are susceptible to vibration, which can affect the contacts and can accelerate wear on

moving parts. Severe vibration can even damage a relay permanently. **Solid-state relays** (discussed in Volume 2) should be used in harsh environments.

resistor

10

OTHER RELATED COMPONENTS

- **potentiometer** (See [Chapter 11](#))

What It Does

A resistor is one of the most fundamental components in electronics. Its purpose is to impede a flow of current and impose a voltage reduction. It consists of two wires or conductors attached at opposite ends or sides of a relatively poor electrical conductor, the resistance of which is measured in ohms, universally represented by the Greek omega symbol, Ω .

Schematic symbols that represent a resistor are shown in [Figure 10-1](#) (Left: The traditional schematic symbol. Right: The more recent European equivalent). The US symbol is still sometimes used in European schematics, and the European symbol is sometimes used in US schematics. Letters K or M indicate that the value shown for the resistor is in thousands of ohms or millions of ohms, respectively. Where these letters are used in Europe, and sometimes in the US, they are substituted for a decimal point. Thus, a 4.7K resistor may be identified as 4K7, a 3.3M resistor may be identified as 3M3, and so on. (The numeric value in [Figure 10-1](#) was chosen arbitrarily.)

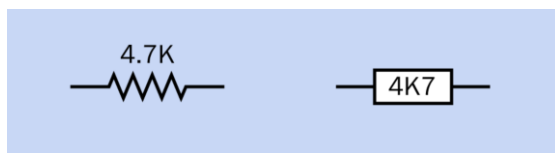


Figure 10-1. Resistor symbols. The left one is more common in the United States, while the right one is widely used in Europe. The 4.7K value was chosen arbitrarily.

A resistor is commonly used for purposes such as limiting the charging rate of a **capacitor**; providing appropriate control voltage to semiconductors such as **bipolar transistors**; protecting **LEDs** or other semiconductors from excessive current; adjusting or limiting the frequency response in an audio circuit (in conjunction with other components); pulling up or pulling down the voltage at the input pin of a digital logic chip; or controlling a voltage at a point in a circuit. In this last application, two resistors may be placed in series to create a [voltage divider](#).

A **potentiometer** may be used instead of a resistor where variable resistance is required.

Sample resistors of various values are shown in [Figure 10-2](#). From top to bottom, their power dissipation ratings are 3W, 1W, 1/2W, 1/4W, 1/4W, and 1/8W. The accuracy (tolerance) of each resistor, from top to bottom, is plus-or-minus 5%, 5%, 5%, 1%, 1%, 5%, and 1%. The beige-colored body of a resistor is often an indication that its tolerance is 5%, while a blue-colored body often indicates a tolerance of 1% or 2%. The blue-bodied resistors and the dark brown resistor contain metal-oxide film elements, while the beige-bodied resistors and the green resistor contain carbon film. For more information on resistor values, see the upcoming [Values](#) section.



Figure 10-2. A range of typical resistors. See text for details.

How It Works

In the process of impeding the flow of current and reducing voltage, a resistor absorbs electrical energy, which it must dissipate as heat. In most modern electronic circuits, the heat dissipation is typically a fraction of a watt.

If R is the resistance in ohms, I is the current flowing through the resistor in amperes, and V is the voltage drop imposed by the resistor (the difference in electrical potential between the two contacts that are attached to it), Ohm's law states:

$$V = I / R$$

This is another way of saying that a resistor of 1Ω will allow a current of 1 amp when the potential difference between the ends of the resistor is 1 volt.

If W is the power in watts dissipated by the resistor, in a DC circuit:

$$W = V * I$$

By substitution in Ohm's law, we can express watts in terms of current and resistance:

$$W = I^2 / R$$

We can also express watts in terms of voltage and resistance:

$$W = V^2 * R$$

These alternates may be useful in situations where you do not know the voltage drop or the current, respectively.

Approximately similar relationships exist when using alternating current, although the power will be a more complex function.

Variants

Axial resistors have two leads that emerge from opposite ends of a usually cylindrical body. *Radial resistors* have parallel leads emerging from one side of the body and are unusual.

Precision resistors are generally defined as having a tolerance of no more than plus-or-minus 1%.

General-purpose resistors are less stable, and their value is less precise.

Power resistors are generally defined as dissipating 1 or 2 watts or more, particularly in power supplies or power amplifiers. They are physically larger and may require *heat sinks* or fan cooling.

Wire-wound resistors are used where the component must withstand substantial heat. A wire-wound resistor often consists of an insulating tube or core that is flat or cylindrical, with multi-

ple turns of resistive wire wrapped around it. The wire is usually a nickel-chromium alloy known as *nichrome* (sometimes written as *Ni-chrome*) and is dipped in a protecting coating.

The heat created by current passing through resistive wire is a potential problem in electronic circuits where temperature must be limited. However, in household appliances such as hair dryers, toaster ovens, and fan heaters, a nichrome element is used specifically to generate heat. Wire-wound resistors are also used in 3D printers to melt plastic (or some other compound) that forms the solid output of the device.

Thick film resistors are sometimes manufactured in a flat, square format. A sample is shown in [Figure 10-3](#), rated to dissipate 10W from its flat surface. The resistance of this component is 1K.

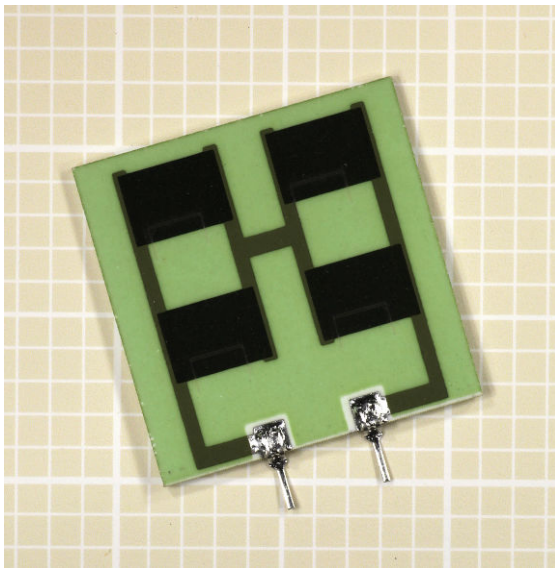


Figure 10-3. A thick-film resistor measuring about 1" square and 0.03" thick.

Surface-mount resistors generally consist of a resistive ink film printed on top of a tablet of aluminum oxide ceramic compound, often approximately 6mm long, known as a 2512 form factor. Each surface-mount resistor has two nickel-

plated terminations coated in solder, which melts when the resistor is attached to the circuit board. The upper surface is coated, usually with black epoxy, to protect the resistive element.

Resistor Array

This is also known as a *resistor network* or *resistor ladder*, and consists of a chip containing multiple equal-valued resistors.

A resistor array in a *single-inline package* (or *SIP*) may have three possible internal configurations: isolated, common bus, and dual terminator. These options are shown at top, center, and bottom, respectively, in [Figure 10-4](#). The isolated variant is commonly available in SIPs with 6, 8, or 10 pins. The common-bus and dual-terminator configurations generally have 8, 9, 10, or 11 pins.

In the isolated configuration, each resistor is electrically independent of the others and is accessed via its own pair of pins. On a common bus, one end of each resistor shares a bus accessed by a single pin, while the other ends of the resistors are accessed by their own separate pins. A dual-terminator configuration is more complex, consisting of pairs of resistors connected between ground and an internal bus, with the midpoint of each resistor pair accessible via a separate pin. The resistor pairs thus function as voltage dividers and are commonly used in emitter-coupled logic circuits that require termination with -2 volts.

A *dual-inline package* (*DIP*) allows a similar range of internal configurations, as shown in [Figure 10-5](#). At top, isolated resistors are commonly available in DIPs with 4, 7, 8, 9, or 10 pins. At center, the common bus configuration is available in DIPs with 8, 14, 16, 18, or 20 pins. At bottom, the dual-terminator configuration usually has 8, 14, 16, 18, or 20 pins.

The external appearance of SIP and DIP resistor arrays is shown in [Figure 10-6](#). From left to right, the packages contain seven 120Ω resistors in

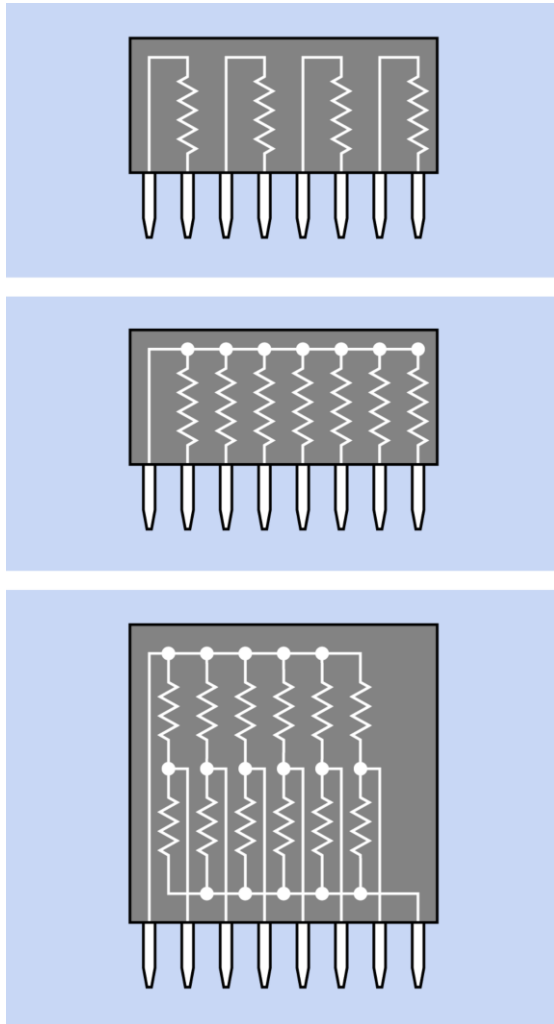


Figure 10-4. Multiple resistors can be embedded in a single-inline package (SIP) in a variety of formats. See text for additional details.

isolated configuration; thirteen 120Ω resistors in bussed configuration; seven $5.6K$ resistors in bussed configuration; and six $1K$ resistors in bussed configuration.

Resistor arrays with isolated or common-bus configurations are a convenient way to reduce the component count in circuits where pullup, pulldown, or terminating resistors are required for multiple chips. The common-bus configuration is also useful in conjunction with a 7-

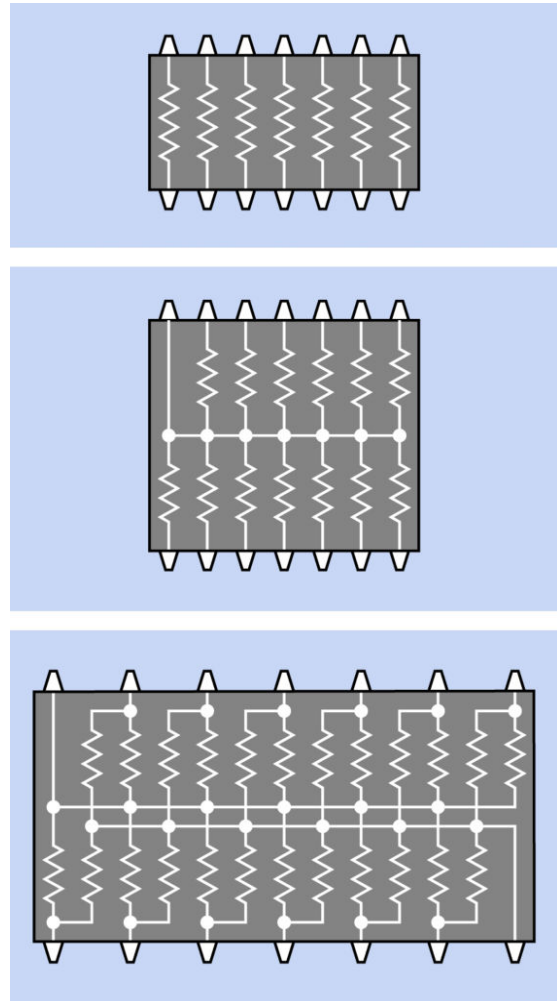


Figure 10-5. Multiple resistors can be obtained embedded in a dual-inline package (DIP). See text for additional details.

segment **LED display**, where each segment must be terminated by a series resistor and all the resistors share a common ground or common voltage source.

Surface-mount chips are available containing a pair of resistors configured as a single voltage divider.

Chips containing multiple *RC circuits* (each consisting of a capacitor and a resistor in series) are available, although uncommon. A package containing a single RC circuit may be sold as a

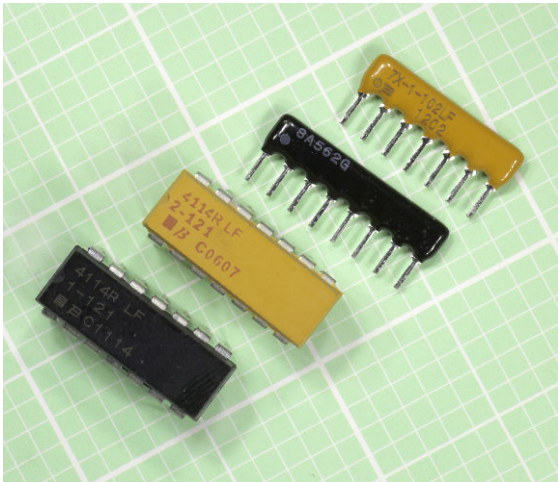


Figure 10-6. Resistor arrays in DIP and SIP packages. See text for values.

snubber to protect contacts in a **switch** or **relay** that switches a large inductive load. More information on snubber circuits is in the **capacitor** entry of this encyclopedia; see “**Snubber**” (page 108).

Values

1 **kilohm**, usually written as 1K, is 1,000Ω. 1 **megohm**, usually written as 1M or 1 meg, is 1,000K. 1 **gigaohm** is 1,000 megs, although the unit is rarely used. Resistances of less than 1Ω are uncommon and are usually expressed as a decimal number followed by the Ω symbol. The term **milliohms** (thousandths of an ohm) is used in special applications. Equivalent resistor values are shown in **Figure 10-7**.

A resistance value remains unchanged in DC and AC circuits, except where the AC reaches an extremely high frequency.

In common electronics applications, resistances usually range from 100Ω to 10M. Power ratings may vary from 1/16 watt to 1000 watts, but usually range from 1/8 watt to 1/2 watt in most electronic circuits (less in surface-mount applications).

Ohms	Kilohms	Megohms
1	0.001	0.000001
10	0.01	0.00001
100	0.1	0.0001
1,000	1	0.001
10,000	10	0.01
100,000	100	0.1
1,000,000	1,000	1

Figure 10-7. Equivalent values in ohms, kilohms, and megohms.

Tolerance

The tolerance, or precision, of a resistor may range from plus-or-minus 0.001% up to plus-or-minus 20%, but is most commonly plus-or-minus 1%, 2%, 5%, or 10%.

The traditional range of resistor values was established when a tolerance of 20% was the norm. The values were spaced to allow minimum risk of a resistor at one end of its tolerance range having the same value as another resistor at the opposite end of its tolerance range. The values were rounded to 10, 15, 22, 33, 47, 68, and 100, as illustrated in **Figure 10-8** where each blue diamond represents the possible range of actual values of a 20% resistor with a theoretical value shown by the white horizontal line at the center of the diamond.

Resistor factors repeat themselves in multiples of 10. Thus, for example, beginning with a resistor of 100Ω, subsequent increasing values will be 150, 220, 330, 470, 680, and 1K, whereas the range of resistors beginning with 1Ω will be 1.5, 2.2, 3.3, 4.7, 6.8, and 10Ω.

Resistance multiplication factors are now expressed as a list of preferred values by the International Electrotechnical Commission (IEC) in their 60063 standard. Intermediate factors have been added to the basic sequence to

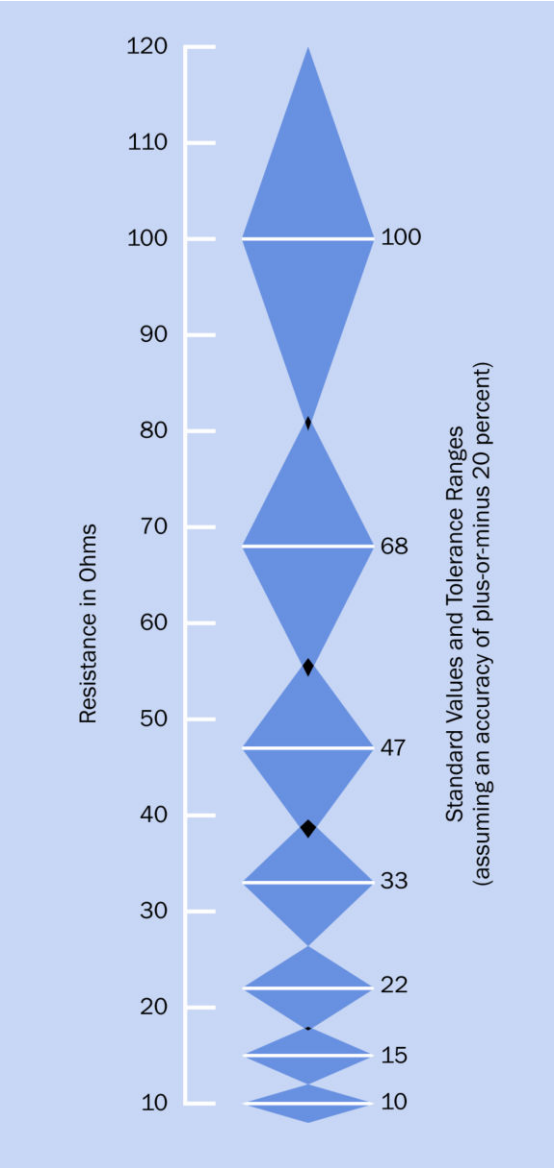


Figure 10-8. Graphical representation of standard resistor values (white lines) established by the International Electrotechnical Commission, showing the acceptable range of actual values (dark blue areas) assuming precision of plus-or-minus 20%. The overlap, if any, between each range and the next is shown in black.

accommodate better tolerances. A table showing resistor values for tolerances of plus-or-minus 20%, 10%, and 5% appears in [Figure 10-9](#). Resistors with a tolerance of 5% have become increasingly common.

	Resistor Tolerances (plus or minus)		
	20%	10%	5%
100	100	100	100
			91
		82	82
			75
68	68	68	68
			62
		56	56
			51
47	47	47	47
			43
		39	39
			36
33	33	33	33
			30
		27	27
			24
22	22	22	22
			20
		18	18
			16
15	15	15	15
			13
		12	12
			11
10	10	10	10

Figure 10-9. Standard values for resistors of different precisions. For resistors outside the range shown, values can be found by multiplying or dividing (repeatedly, if necessary) by a factor of 10.

The IEC has established 3-digit preferred values for resistors with values accurate to plus-or-minus 0.5%.

Because many capacitors still have a tolerance no better than 20%, their values also conform with

the old original set of resistance values, although the units are expressed in farads or fractions of a farad. See the **capacitors** entry in this encyclopedia for additional information.

Value Coding

Through-hole axial resistors are traditionally printed with a sequence of three colored bands to express the value of the component, each of the first two bands representing a digit from 0 through 9, while the third band indicates the decimal multiplier (the number of zeroes, from 0 to 9, which should be appended to the digits). A fourth band of silver or gold indicates 10% or 5% tolerance respectively. No fourth band would indicate 20% tolerance, although this has become very rare.

Many resistors now have five color bands, to enable the representation of intermediate or fractional values. In this scheme, the first three bands have numeric values (using the same color system as before) while the fourth band is the multiplier. A fifth band, at the opposite end of the resistor, indicates its tolerance.

In [Figure 10-10](#) the numeric or multiplier value of each color is shown as a “spectrum” at the top of the figure. The tolerance, or precision of a resistor, expressed as a plus-or-minus percentage, is shown using silver, gold, and various colors, at the bottom of the figure.

Two sample resistors are shown. The upper one has a value of 1K, indicated by the brown and black bands on the left (representing numeral 1 followed by a numeral 0) and the third red band (indicating two additional zeroes). The gold band at right indicates a precision of 5%. The lower one has a value of 1.05K, indicated by the brown, black, and green bands on the left (representing numeral 1 followed by numeral 0 followed by a numeral 5) and the fourth band brown (indicating one additional zero). The brown band at right indicates a precision of 1%.

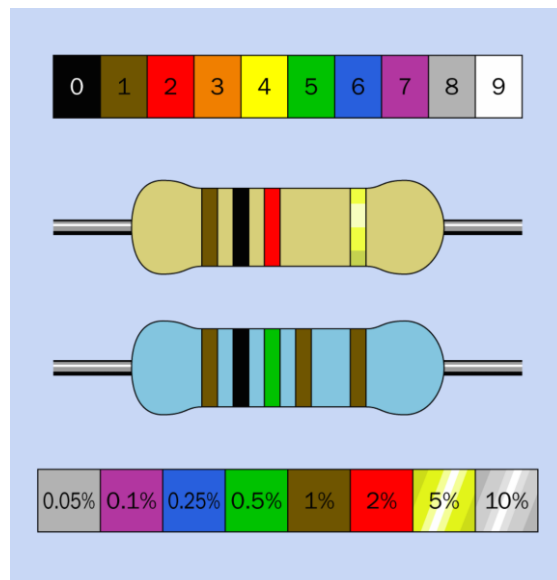


Figure 10-10. Color coding of through-hole resistors. See text for details.

In extremely old equipment, resistors may be coded with the body-tip-dot scheme, in which the body color represents the initial digit, the end color represents the second digit, and a dot represents the multiplier. The numeric identities of the colors is the same as in the current color scheme.

In all modern schemes, the three or four bands that show the resistance value are spaced close together, while a larger gap separates them from the band that shows the tolerance. The resistor value should be read while holding the resistor so that the group of closely-spaced numeric bands is on the left.

Confusingly, some resistors may be found where the first three bands define the value, using the old three-band convention; the fourth band indicates tolerance; and a fifth band at the opposite end of the component indicates reliability. However, this color scheme is uncommon.

Other color-coding conventions may be found in special applications, such as military equipment.

It is common for through-hole carbon-film resistors to have a beige body color, while through-hole metal-film resistors often have a blue body color. However, in relatively rare instances, a blue body color may also indicate a *fusible* resistor (designed to burn out harmlessly like a fuse, if it is overloaded) while a white body may indicate a *non-flammable* resistor. Use caution when replacing these special types.

Some modern resistors may have their values printed on them numerically. Surface-mount resistors also have digits printed on them, but they are a code, not a direct representation of resistance. The last digit indicates the number of zeroes in the resistor value, while the preceding two or three numbers define the value itself. Letter R is used to indicate a decimal point. Thus a 3R3 surface-mount resistor has a value of 3.3Ω, while 330 would indicate 33Ω, and 332 indicates 3,300Ω. A 2152 surface-mount resistor would have a value of 21,500Ω.

A surface-mount resistor with a single zero printed on it is a *zero ohm* component that has the same function as a **jumper** wire. It is used for convenience, as it is easily inserted by automated production-line equipment. It functions merely as a bridge between traces on the circuit board.

When resistor values are printed on paper in schematics, poor reproduction may result in omission of decimal points or introduction of specks that look like decimal points. Europeans have addressed this issue by using the letter as a substitute for a decimal point so that a 5.6K resistor will be shown as 5K6, or a 3.3M resistor will be shown as 3M3. This practice is followed infrequently in the United States.

Stability

This term describes the ability of a resistor to maintain an accurate value despite factors such as temperature, humidity, vibration, load cycling, current, and voltage. The *temperature coefficient* of a resistor (often referred to as T_{cr} or T_c , not to be confused with the time constant of a charging capacitor) is expressed in parts per million

change in resistance for each degree centigrade deviation from room temperature (usually assumed to be 25 degrees Centigrade). T_c may be a positive or a negative value.

The *voltage coefficient* of resistance—often expressed as V_c —describes the change of a resistor's value that may occur as a function of changes in voltage. This is usually significant only where the resistive element is carbon-based. If V_1 is the rated voltage of the resistor, R_1 is its rated resistance at that voltage, V_2 is 10% of the rated voltage, and R_2 is the actual resistance at that voltage, the voltage coefficient, V_c , is given by this formula:

$$V_c = (100 * (R_1 - R_2)) / (R_2 * (V_1 - V_2))$$

Materials

Resistors are formed from a variety of materials.

Carbon composite. Particles of carbon are mixed with a binder. The density of the carbon determines the end-to-end resistance, which typically ranges from 5Ω to 10M. The disadvantages of this system are low precision (a 10% tolerance is common), relatively high voltage coefficient of resistance, and introduction of noise in sensitive circuits. However, carbon-composite resistors have low inductance and are relatively tolerant of overload conditions.

Carbon film. A cheap and popular type, made by coating a ceramic substrate with a film of carbon compound. They are available in both through-hole and surface-mount formats. The range of resistor values is comparable with carbon-composite types, but the precision is increased, typically to 5%, by cutting a spiral groove in the carbon-compound coating during the manufacturing process. The carbon film suffers the same disadvantages of carbon composite resistors, but to a lesser extent. Carbon film resistors generally should not be substituted for metal film resistors in applications where accuracy is important.

Metal film. A metallic film is deposited on a ceramic substrate, and has generally superior char-

acteristics to carbon-film resistors. During manufacture, a groove may be cut in the metal film to adjust the end-to-end resistance. This may cause the resistor to have higher inductance than carbon-composite types, though it has lower noise. Tolerances of 5%, 2%, and 1% are available. This type of resistor was originally more expensive than carbon-film equivalents, but the difference is now fractional. They are available in both through-hole and surface-mount formats. They are available in lower-wattage variants (1/8 watt is common).

Thick-film resistors are spray-coated, whereas thin-film resistors are sputtered nichrome. Thin-films enjoy a flatter temperature coefficient and are typically used in environments that have a huge operational temperature range, such as satellites.

Bulk metal foil. The type of foil used in metal film resistors is applied to a ceramic wafer and etched to achieve the desired overall resistance. Typically these resistors have axial leads. They can be extremely accurate and stable, but have a limited maximum resistance.

Precision wire-wound. Formerly used in applications requiring great accuracy, but now largely replaced by precision metal foil.

Power wire-wound. Generally used when 1 or 2 watts or more power dissipation is required. Resistive wire is wrapped around a core that is often ceramic. This can cause the resistor to be referred to, inaccurately, as “ceramic.” The core may alternatively be fiberglass or some other electrically insulating compound that actively sinks heat. The component is either dipped (typically in vitreous enamel or cement) or is mounted in an aluminum shell that can be clamped to a heat sink. It almost always has the ohm value printed on it in plain numerals (not codes).

Two typical wire-wound resistors are shown in Figure 10-11. The upper resistor is rated at 12W and 180Ω while the lower resistor is rated at 13W and 15K.



Figure 10-11. Two wire-wound resistors of greatly differing resistance but similar power dissipation capability.

A larger wire-wound resistor is shown in Figure 10-12, rated for 25W and 10Ω.



Figure 10-12. A large wirewound resistor rated to dissipate 25W.

In Figure 10-13, two resistors encapsulated in cement coatings are shown with the coatings removed to expose the elements. At left is a 1.5Ω 5W resistor, which uses a wire-wound element. At right is a very low-value 0.03Ω 10W resistor.

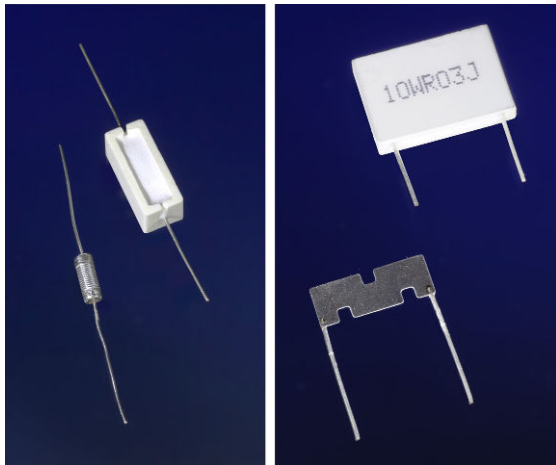


Figure 10-13. Two low-value resistors with their cement coatings removed to show the resistive elements.

In [Figure 10-14](#), the resistor at right has an exposed 30Ω element while the resistor at left is rated 10W and 6.5Ω , enclosed in an anodized aluminum shell to promote heat dissipation.

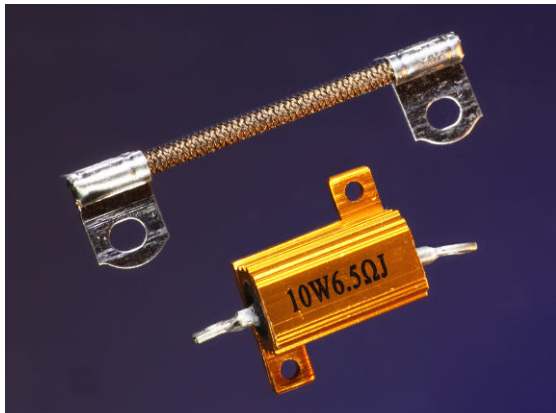


Figure 10-14. A 30Ω resistor (right) and 6.5Ω resistor (left).

In power resistors, heat dissipation becomes an important consideration. If other factors (such as voltage) remain the same, a lower-value resistor will tend to pass more current than a higher-value resistor, and heat dissipation is proportional to the square of the current. Therefore power wire-wound resistors are more likely to be need-

ed where low resistance values are required. Their coiled-wire format creates significant inductance, making them unsuitable to pass high frequencies or pulses.

How to Use it

Some of the most common applications for a resistor are listed here.

In Series with LED

To protect an LED from damage caused by excessive current, a [series resistor](#) is chosen to allow a current that does not exceed the manufacturer's specification. In the case of a single through-hole LED (often referred to as an [indicator](#)), the forward current is often limited to around 20mA, and the value of the resistor will depend on the voltage being used. (See [Figure 10-15](#).)

When using high-output LEDs (which may contain multiple elements in a single 5mm or 10mm package), or LED arrays that are now being used for domestic lighting, the acceptable current may be much greater, and the LED unit may contain its own current-limiting electronics. A data-sheet should be consulted for details.

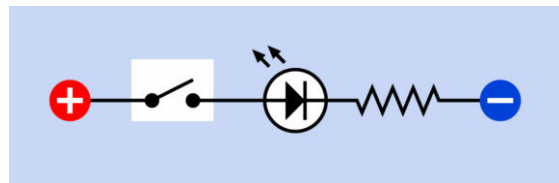


Figure 10-15. A series resistor is necessary to limit the current that passes through an LED.

Current Limiting with a Transistor

In [Figure 10-16](#), a transistor is switching or amplifying current flowing from B to C. A resistor is used to protect the base of the transistor from excessive current flowing from point A. Resistors are also commonly used to prevent excessive current from flowing between B and C.

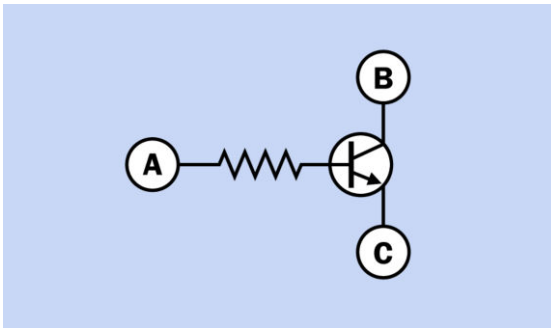


Figure 10-16. A resistor is typically necessary to protect the base of a transistor from excessive current.

Pullup and Pulldown Resistors

When a mechanical switch or pushbutton is attached to the input of a logic chip or **microcontroller**, a *pullup* or *pulldown* resistor is used, applying positive voltage or grounding the pin, respectively, to prevent it from “floating” in an indeterminate state when the switch is open. In [Figure 10-17](#), the upper schematic shows a pulldown resistor, whereas the lower schematic shows a pullup resistor. A common value for either of them is 10K. When the pushbutton is pressed, its direct connection to positive voltage or to ground easily overwhelms the effect of the resistor. The choice of pullup or pulldown resistor may depend on the type of chip being used.

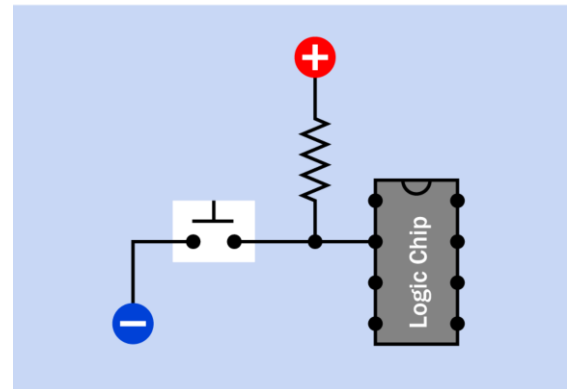
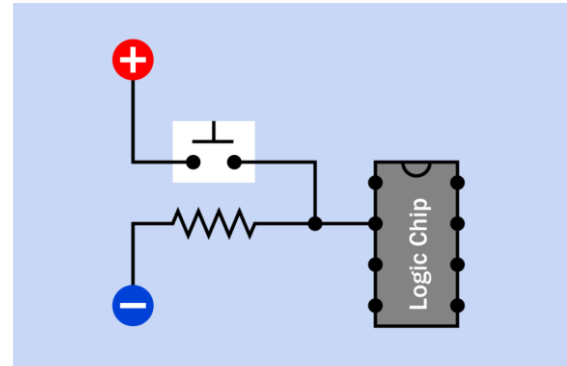


Figure 10-17. A pulldown resistor (top) or pullup resistor (bottom) prevents an input pin on a logic chip or microcontroller from “floating” in an indeterminate state when the button is not being pressed.

Audio Tone Control

A resistor-capacitor combination can limit the high-frequency in a simple audio tone-control circuit, as shown in [Figure 10-18](#). Beneath a signal travelling from A to B, a resistor is placed in series with a capacitor that passes high frequencies to ground. This is known as a low-pass filter.

RC Network

A resistor will adjust the charge/discharge time when placed in series with a capacitor, as in [Figure 10-19](#). When the switch closes, the resistor limits the rate at which the capacitor will charge itself from the power supply. Because a capacitor has an ideally infinite resistance to DC current, the voltage measured at point A will rise until it

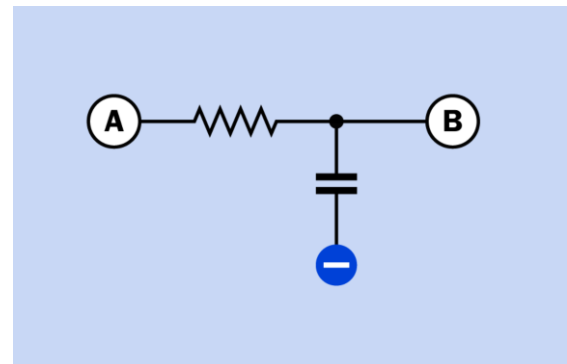


Figure 10-18. This configuration may be used to remove high frequencies from an audio signal. It is known as a low-pass filter because low frequencies are passed from A to B.

is close to the supply voltage. This is often referred to as an **RC (resistor-capacitor)** network and is discussed in greater detail in the **capacitor** section of this encyclopedia.

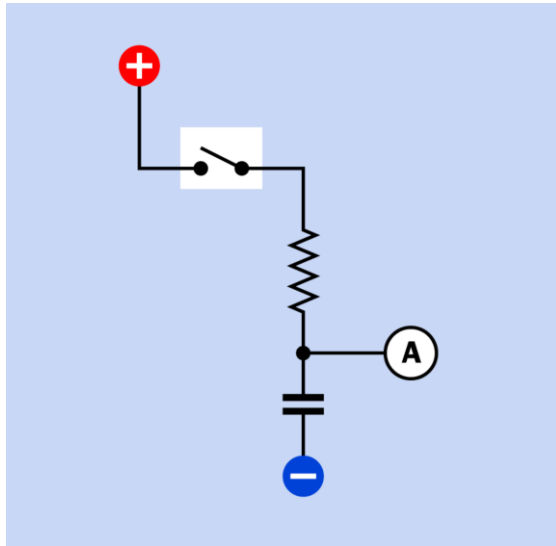


Figure 10-19. In an **RC (resistor-capacitor)** network, a resistor limits the rate of increase in potential of the capacitor, measured at A, when the switch is closed.

Voltage Divider

Two resistors may be used to create a voltage divider (see **Figure 10-20**). If V_{in} is the supply voltage, the output voltage, V_{out} , measured at point A, is found by the formula:

$$V_{out} = V_{in} * (R2 / (R1 + R2))$$

In reality, the actual value of V_{out} is likely to be affected by how heavily the output is loaded.

If the output node has a high impedance, such as the input to a logic chip or comparator, it will be more susceptible to electrical noise, and lower-value resistors may be needed in the voltage divider to maintain a higher current flow and maintain stability in the attached device.

Resistors in Series

If resistors in series have values $R1, R2, R3 \dots$ the total resistance, R , is found by summing the individual resistances:

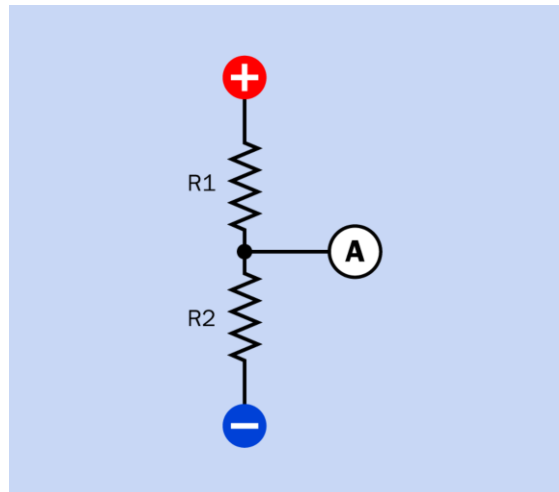


Figure 10-20. In a DC circuit, a pair of resistors may be placed in series to function as a voltage divider. The voltage measured at A will be lower than the supply voltage, but above ground potential.

$$R = R1 + R2 + R3 \dots$$

The current through each of the resistors will be the same, whereas the voltage across each of them will vary proportionately with its resistance. If the supply voltage across the series of resistors is V_S , and the total of all the resistor values is R_T , and the resistance of one resistor is $R1$, the voltage across that resistor, $V1$, will be given by the formula:

$$V1 = V_S * (R1 / R_T)$$

Resistors in Parallel

Where two or more resistors ($R1, R2, R3 \dots$) are wired in parallel, their total resistance, R , is found from the formula:

$$1/R = (1/R1) + (1/R2) + (1/R3) \dots$$

Suppose that $R1, R2, R3 \dots$ all have the same individual resistance, represented by R_I , and the number of resistors is N . Their total resistance, R_T , when wired in parallel, will be:

$$R_T = R_I / N$$

If each resistor has an equal resistance and also has an equal individual rating in watts (represented by WI), the total wattage (WT) that they can handle when wired in parallel to share the power will be:

$$WT = WI * N$$

Therefore, if an application requires high-wattage resistors, multiple lower-wattage, higher-value resistors may be substituted if they are wired in parallel—and may even be cheaper than a single high-wattage wire-wound resistor. For example, if a 5W, 50Ω resistor is specified, 10 resistors can be substituted, each rated at 0.5W and 500Ω. Bear in mind that if they are tightly bundled, this will interfere with heat dissipation.

What Can Go Wrong

Heat

Resistors are probably the most robust of all electronic components, with high reliability and a long life. It is difficult to damage a resistor by overheating it with a soldering iron.

The wattage rating of a resistor does not necessarily mean that it should be used to dissipate that amount of power on a constant basis. Small resistors (1/4 watt or less) can overheat just as easily as big ones. Generally speaking, it is safe practice not to exceed 75% of a resistor's power rating on a constant basis.

Overheating is predictably more of a problem for power resistors, where provision must be made for heat dissipation. Issues such as component crowding should be considered when deciding how big a heat sink to use and how much ventilation. Some power resistors may function reliably at temperatures as high as 250 degrees Centigrade, but components near them are likely to be less tolerant and plastic enclosures may soften or melt.

Noise

The electrical *noise* introduced by a resistor in a circuit will vary according to the composition of

the resistor, but for any given component, it will be proportional to voltage and current. Low-noise circuits (such as those at the input stage of a high-gain amplifier) should use low-wattage resistors at a low voltage where possible.

Inductance

The coiled wire of a wire-wound resistor will be significantly inductive at low frequencies. This is known as *parasitic inductance*. It will also have a resonant frequency. This type of resistor is unsuitable for applications where frequency exceeds 50KHz.

Inaccuracy

When using resistors with 10% tolerance, imprecise values may cause greater problems in some applications than in others. In a voltage divider, for instance, if one resistor happens to be at the high end of its tolerance range while the other happens to be at the low end, the voltage obtained at the intersection of the resistors will vary from its expected value. Using the schematic shown in [Figure 10-20](#), if R1 is rated for 1K and R2 is rated for 5K, and the power supply is rated at 12VDC, the voltage at point A should be:

$$V = 12 * ((5 / (5 + 1))) = 10$$

However, if R1 has an actual value of 1.1K and R2 has an actual value of 4.5K, the actual voltage obtained at point A will be:

$$V = 12 * ((4.5 / (4.5 + 1.1))) = 9.6$$

If the resistors are at opposite ends of their respective tolerance ranges, so that R1 has an actual value of 900Ω while the lower resistor has an actual value of 5.5K, the actual voltage obtained will be:

$$V = 12 * ((5.5 / (5.5 + 0.9))) = 10.3$$

The situation becomes worse if the two resistors are chosen to be of equal value, to provide half of the supply voltage (6 volts, in this example) at their intersection. If two 5K resistors are used, and the upper one is actually 4.5K while the lower one is 5.5K, the actual voltage will be:

$$V = 12 * ((5.5 / (4.5 + 5.5))) = 6.6$$

Whether this variation is significant will depend on the particular circuit in which the voltage divider is being used.

Common through-hole resistors may occasionally turn out to have values that are outside their specified tolerance range, as a result of poor manufacturing processes. Checking each resistor with a meter before placing it in a circuit should be a standard procedure.

When measuring the voltage drop introduced by a resistor in an active circuit, the meter has its own internal resistance that will take a proportion of the current. This is known as *meter loading* and will result in an artificially low reading for

the potential difference between the ends of a resistor. This problem becomes significant only when dealing with resistors that have a high value (such as 1M), comparable with the internal resistance of the meter (likely to be 10M or more).

Wrong Values

When resistors are sorted into small bins by the user, errors may be made, and different values may be mixed together. This is another reason for checking the values of components before using them. Identification errors may be nontrivial and easily overlooked: the visible difference between a 1 megohm resistor and a 100Ω resistor is just one thin color band.

potentiometer

Also known as a *variable resistor*; may be substituted for a *rheostat*.

OTHER RELATED COMPONENTS

- **rotational encoder** (see [Chapter 8](#))
- **resistor** (see [Chapter 10](#))

What It Does

When a voltage is applied across a potentiometer, it can deliver a variable fraction of that voltage. It is often used to adjust sensitivity, balance, input, or output, especially in audio equipment and sensors such as motion detectors.

A potentiometer can also be used to insert a variable resistance in a circuit, in which case it should really be referred to as a *variable resistor*, although most people will still call it a potentiometer.

It can be used to adjust the power supplied to a circuit, in which case it is properly known as a *rheostat*, although this term is becoming obsolete. Massive rheostats were once used for purposes such as dimming theatrical lighting, but solid-state components have taken their place in most high-wattage applications.

A full-size, classic-style potentiometer is shown in [Figure 11-1](#).

Schematic symbols for a potentiometer and other associated components are shown in [Figure 11-2](#), with American versions on the left and European versions on the right in each case. The symbols for a potentiometer are at the top. The correct symbols for a variable resistor or rheostat are shown at center, although a potentiometer symbol may often be used instead. A



Figure 11-1. A generic or classic-style potentiometer, approximately one inch in diameter.

preset variable resistor is shown at the bottom, often referred to as a *trimmer* or *Trimpot*. In these examples, each has an arbitrary rated resistance of 4,700Ω. Note the European substitution of K for a decimal point.

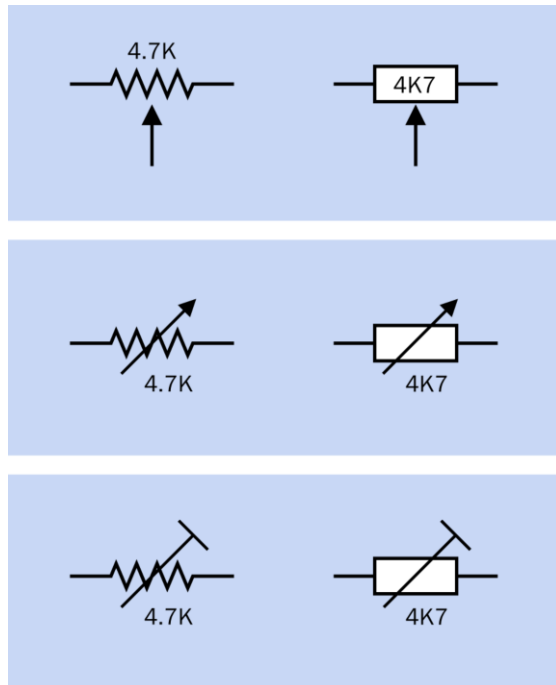


Figure 11-2. American (left) and European (right) symbols for a potentiometer, a rheostat, and a trimmer potentiometer, reading from top to bottom. The 4.7K value was chosen arbitrarily.

How It Works

A potentiometer has three terminals. The outer pair connect with the opposite ends of an internal resistive element, such as a strip of conductive plastic, sometimes known as the *track*. The third center terminal connects internally with a contact known as the *wiper* (or rarely, the *pick-off*), which touches the strip and can be moved from one end of it to the other by turning a shaft or screw, or by moving a slider.

If an electrical potential is applied between opposite ends of the resistive element, the voltage “picked off” by the wiper will vary as it moves. In this mode, the potentiometer works as a resistive voltage divider. For example, in a potentiometer with a linear taper (see “Variants,” coming up), if you attach the negative side of a 12V battery to the right-hand end terminal and the positive side to the left-hand end terminal, you will find an 8V

potential at the center terminal when the potentiometer has rotated clockwise through one-third of its range. In Figure 11-3, the base of the shaft (shown in black) is attached to an arm (shown in green) that moves a wiper (orange) along a resistive element (brown). The voltages shown assume that the resistive element has a linear taper and will vary slightly depending on wire resistance and other factors.

Because a potentiometer imposes a voltage reduction, it also reduces current flowing through it, and therefore creates waste heat which must be dissipated. In an application such as an audio circuit, small currents and low voltages generate negligible heat. If a potentiometer is used for heavier applications, it must be appropriately rated to handle the wattage and must be vented to allow heat to disperse.

To use a potentiometer as a *variable resistor* or *rheostat*, one of its end terminals may be tied to the center terminal. If the unused end terminal is left unconnected, this raises the risk of picking up stray voltages or “noise” in sensitive circuits. In Figure 11-4, a potentiometer is shown adjusting a series resistance for an LED for demonstration purposes. More typically, a trimmer would be used in this kind of application, since a user is unlikely to need to reset it.

Variants

Linear and Log Taper

If the resistive element in a potentiometer is of constant width and thickness, the electrical potential at the wiper will change in ratio with the rotation of the wiper and shaft (or with movement of a slider). This type of potentiometer is said to have a linear taper even though its element does not actually taper.

For audio applications, because human hearing responds nonlinearly to sound pressure, a potentiometer that has a linear taper may seem to have a very slow action at one end of its scale and an abrupt effect at the other. This problem used to be solved with a non-uniform or tapered re-

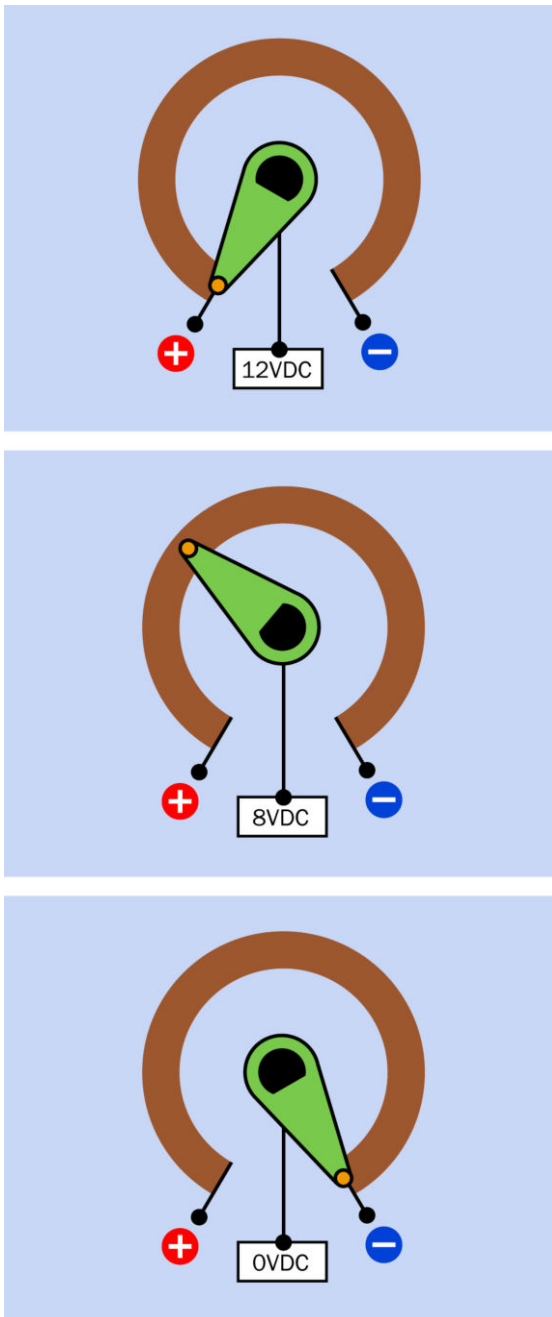


Figure 11-3. Inside a potentiometer. See text for details.

sistive element. More recently, a combination of resistive elements has been used as a cheaper option. Such a potentiometer is said to have an audio taper or a log taper (since the resistance

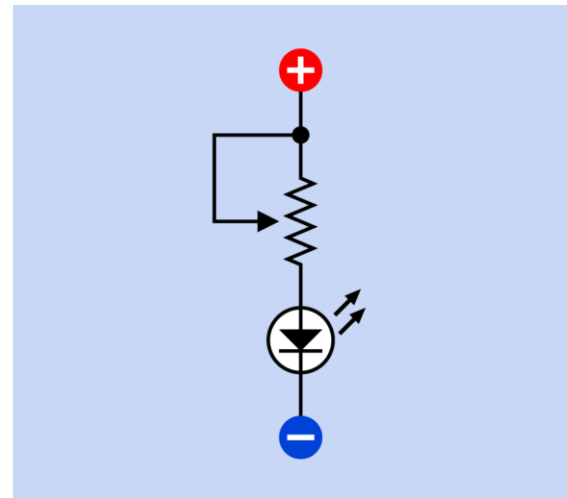


Figure 11-4. A potentiometer can be used to adjust a series resistance, as shown in this schematic. Tying the wiper to one of the end terminals reduces the risk of picking up electrical noise.

may vary as a logarithm of the angle of rotation). A reverse audio taper or antilog taper varies in the opposite direction, but this type has become very uncommon.

Classic-style Potentiometer

This consists of a sealed circular can, usually between 0.5" and 1" in diameter, containing a resistive strip that is shaped as a segment of a circle. A typical example is shown in [Figure 11-1](#), although miniaturized versions have become more common. A shaft mounted on the can turns the internal wiper that presses against the strip. For panel-mount applications, a threaded bushing at the base of the shaft is inserted through a hole in the front panel of the electronics enclosure, and a nut is tightened on the bushing to hold the potentiometer in place. Often there is also a small offset index pin that, when paired with a corresponding front panel hole, will keep the pot from spinning freely.

Many modern potentiometers are miniaturized, and may be packaged in a box-shaped plastic

enclosure rather than a circular can. Their power ratings are likely to be lower, but their principle of operation is unchanged. Two variants are shown in [Figure 11-5](#).

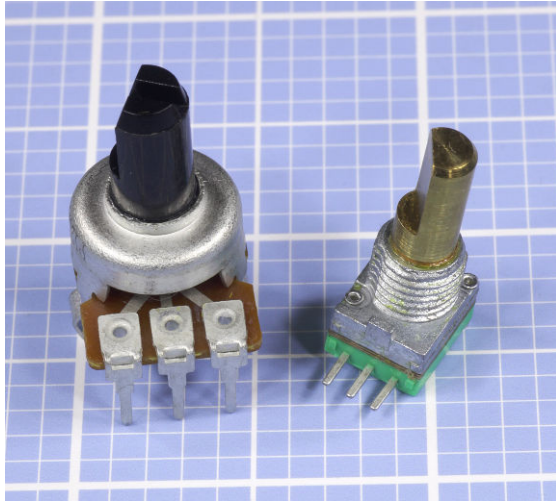


Figure 11-5. Two modern miniaturized potentiometers. At left: 5K. At right: 10K. Both are rated to dissipate up to 50mW.

The three terminals on the outside of a potentiometer may be solder lugs, screw terminals, or pins for direct mounting on a circuit board. The pins may be straight or angled at 90 degrees.

The resistive element may use carbon film, plastic, cermet (a ceramic-metal mixture), or resistive wire wound around an insulator. Carbon-film potentiometers are generally the cheapest, whereas wire-wound potentiometers are generally the most expensive.

Wire-wound potentiometers may handle more power than the other variants, but as the wiper makes a transition from one turn of the internal wire element to the next, the output will tend to change in discrete steps instead of varying more smoothly.

In a potentiometer with detents, typically a spring-loaded lever in contact with notched internal wheel causes the shaft to turn in discrete steps that create a stepped output even if the resistive element is continuous.

The shaft may be made of metal or plastic, with its length and width varying from one component to another. A control knob can be fitted to the end of the shaft. Some control knobs are push-on, others have a set screw to secure them. Shafts may be splined and split, or round and smooth, or round with a flat surface that matches the shape of a socket in a control knob and reduces the risk of a knob becoming loose and turning freely. Some shafts have a slotted tip to enable screwdriver adjustment.

Some shaft options for full-size potentiometers are shown in [Figure 11-6](#).



Figure 11-6. Three shaft options for potentiometers.

Multiple-Turn Potentiometer

To achieve greater precision, a track inside a potentiometer may be manufactured in the form of a helix, allowing the wiper to make multiple turns on its journey from one end of the track to the other. Such multiple-turn potentiometers typically allow 3, 5, or 10 turns to move the wiper from end to end. Other multiple-turn potentiometers may use a screw thread that advances a wiper along a linear or circular track. The latter is com-

parable with a trimmer where multiple turns of a screwdriver are used to rotate a worm gear that rotates a wiper between opposite ends of a circular track.

Ganged Potentiometer

Two (or rarely, more) potentiometers can be stacked or combined so that their resistive elements and wipers share the same shaft but can use different voltages or have different taper. Each resistance-wiper assembly is known as a *cup*, and the potentiometers are said to be *ganged*.

Flat ganged potentiometers combine two resistive elements in one enclosure. Some *dual ganged* potentiometers are concentric, meaning that the pots are controlled separately by two shafts, one inside the other. Suitable concentric knobs must be used. You are unlikely to find these potentiometers sold as components in limited quantities.

Switched Potentiometer

In this variant, when the shaft is turned clockwise from an initial position that is fully counter-clockwise, it flips an internal switch connected to external terminals. This can be used to power-up associated components (for example, an audio amplifier). Alternatively, a switch inside a potentiometer may be configured so that it is activated by pulling or pushing the shaft.

Slider Potentiometer

Also known as a *slide potentiometer*. This uses a straight resistive strip and a wiper that is moved to and fro linearly by a tab or lug fitted with a plastic knob or finger-grip. Sliders are still found on some audio equipment. The principle of operation, and the number of terminals, are identical to the classic-style potentiometer. Sliders typically have solder tabs or PC pins. In [Figure 11-7](#), the large one is about 3.5" long, designed for mounting behind a panel that has a slot to allow the sliding lug to poke through. Threaded holes at either end will accept screws to fix the slider behind the panel. A removable plastic finger-grip

(sold separately, in a variety of styles) has been pushed into place. Solder tabs underneath the slider are hidden in this photo. The smaller slider is designed for through-hole mounting on a circuit board.

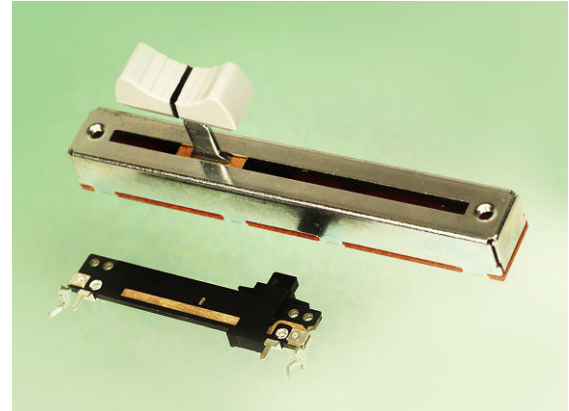


Figure 11-7. Slider potentiometers.

Trimmer Potentiometer

Often referred to as Trimpots, this is actually a proprietary brand name of Bourns. They are usually mounted directly on circuit boards to allow fine adjustment or trimming during manufacturing and testing to compensate for variations in other components. Trimmers may be single-turn or multi-turn, the latter containing a worm gear that engages with another gear to which the wiper is attached. Trimmers always have linear taper. They may be designed for screwdriver adjustment or may have a small knurled shaft, a thumb wheel, or a knob. They are not usually accessible by the end user of the equipment, and their setting may be sealed or fixed when the equipment is assembled. In [Figure 11-8](#), the beige Spectrol trimmer is a single-turn design, whereas the blue trimmer is multi-turn. A worm gear inside the package, beneath the screw head, engages with an interior gear wheel that rotates the wiper.

In [Figure 11-9](#), a 2K trimmer potentiometer has a knurled dial attached to allow easy finger adjustment, although the dial also contains a slot for a flat-blade screwdriver.

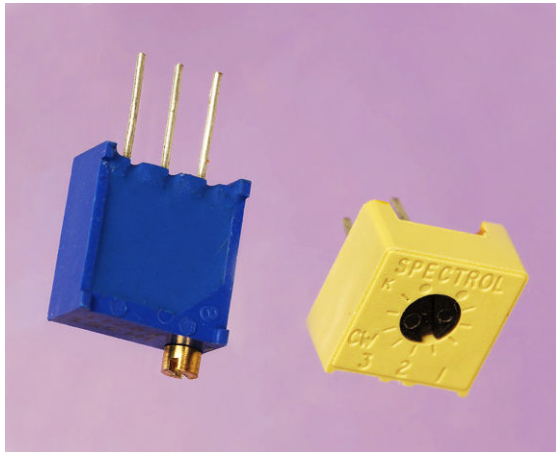


Figure 11-8. Like most trimmers, these are designed for through-hole mounting on a circuit board.

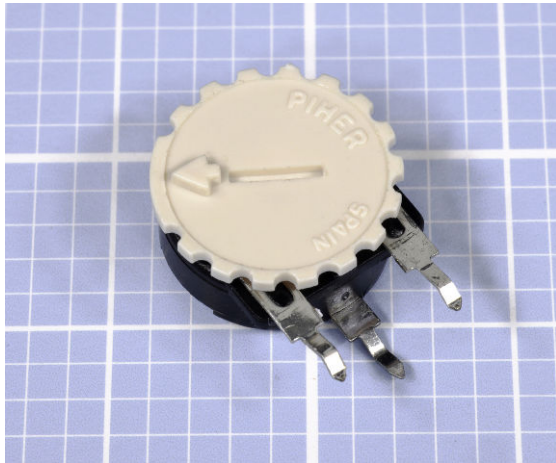


Figure 11-9. A trimmer potentiometer with a knurled dial to facilitate finger adjustment.

How to Use it

The classic-style potentiometer was once used universally to control volume, bass, and treble on audio equipment but has been replaced increasingly by digital input devices such as **tactile switches** (see “Tactile Switch” (page 34)) or **rotational encoders** (see Chapter 8), which are more reliable and may be cheaper, especially when assembly costs are considered.

Potentiometers are widely used in lamp dimmers and on cooking stoves (see Figure 11-10). In these applications, a solid-state switching device such as a **triac** (described in Volume 2) does the actual work of moderating the power to the lamp or the stove by interrupting it very rapidly. The potentiometer adjusts the duty cycle of the power interruptions. This system wastes far less power than if the potentiometer controlled the lighting or heating element directly as a rheostat. Since less power is involved, the potentiometer can be small and cheap, and will not generate significant heat.

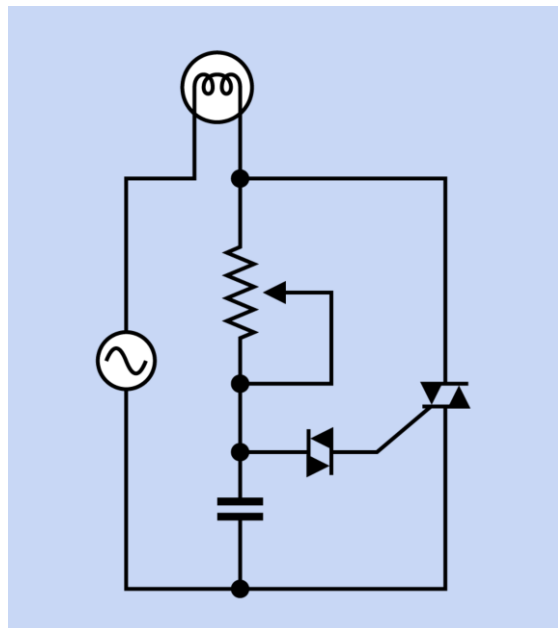


Figure 11-10. Typical usage of a potentiometer in conjunction with a diac, triac, and capacitor to control the brightness of an incandescent bulb, using an AC power supply. Diacs and triacs are discussed in Volume 2.

Because true logarithmic potentiometers have become decreasingly common, a linear potentiometer in conjunction with a fixed resistor can be used as a substitute, to control audio input. See Figure 11-11.

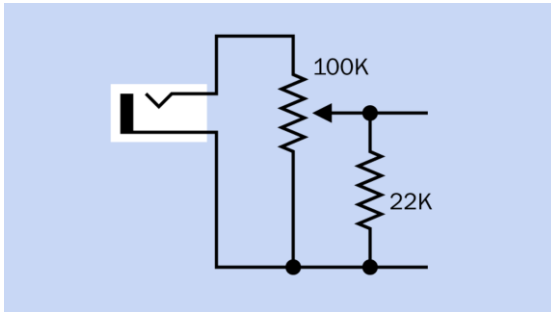


Figure 11-11. In this circuit, a 100K linear potentiometer is used in conjunction with a 22K resistor to create an approximately logarithmic volume control for an audio system with input coming from a mono jack socket at left.

A potentiometer may be used to match a sensor or analog input device to an analog-digital converter, or it can calibrate a device such as a temperature or motion sensor.

What Can Go Wrong

Wear and Tear

Since classic-style potentiometers are electro-mechanical devices, their performance will deteriorate as one part rubs against another. The long open slot of a slider potentiometer makes it especially vulnerable to contamination with dirt, water, or grease. Contact-cleaner solvent, lubricant-carrying sprays, or pressurized “duster” gas may be squirted into a potentiometer to try to extend its life. Carbon-film potentiometers are the least durable and in audio applications will eventually create a “scratchy” sound when they are turned, as the resistive element deteriorates.

If the wiper deteriorates to the point where it no longer makes electrical contact with the track, and if the potentiometer is being used as a variable resistor, two failure modes are possible, shown in Figure 11-12. Clearly the right-hand schematic is a better outcome. This is an argument for always tying the wiper to the “unused” end of the track.

If you are designing a circuit board that will go through a production process, temperature variations during wave soldering, and subsequent

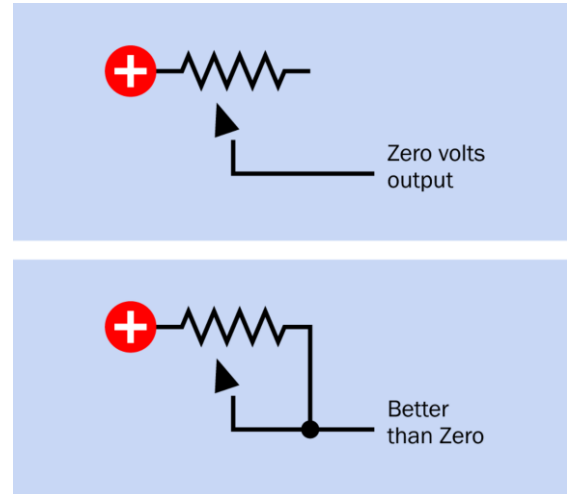


Figure 11-12. If the wiper of a potentiometer breaks (indicated by the loose arrow head) as a result of wear and tear, and the potentiometer is being used as a variable resistor, the voltage from it will drop to zero (top schematic) unless the wiper has been tied to one end of the track (bottom schematic).

washing to remove flux residues, create hostile conditions for potentiometers, especially sliders where the internal parts are easily contaminated. It will be safer to hand-mount potentiometers after the automated process.

Knobs that Don't Fit

Control knobs are almost always sold separately from potentiometers. Make sure the shaft of the potentiometer (which may be round, round-with-flat, or knurled) matches the knob of your choice. Note that some shaft diameters are expressed in inches, while others are metric.

Nuts that Get Lost

For panel-mounted potentiometers, a nut that fits the thread on the bushing is almost always included with the potentiometer; an additional nut and lock washer may also be supplied. Because there is no standardization of threads on potentiometers, if you lose a nut, you may have some difficulty finding an exact replacement.

A Shaft that Isn't Long Enough

When choosing a shaft length, if in doubt, buy a potentiometer with a long shaft that you can cut to the desired length.

Sliders with No Finger Grip

Slider potentiometers are often sold without a knob or plastic finger-grip, which must be ordered separately and may be available in different styles. The finger-grip usually push-fits onto the metal or plastic tab or lug that moves the slider to and fro.

Too Big to Fit

Check the manufacturer's datasheet if you need to know the physical size of the potentiometer. Photographs may be misleading, as a traditional-style potentiometer that is 0.5" in diameter looks much the same as one that is 1" in diameter. High-wattage potentiometers will be more costly and physically large (2 to 3 inches in diameter). See Figure 11-13.

Overheating

Be sure to leave sufficient air space around a high-wattage potentiometer. Carefully calculate the maximum voltage drop and current that you may be using, and choose a component that is appropriately rated. Note that if you use the potentiometer as a rheostat, it will have to handle more current when its wiper moves to reduce its resistance. For example, if 12VDC are applied through a 10-ohm rheostat to a component that has a resistance of 20 ohms, current in the circuit will vary from 0.4 amps to 0.6 amps depending on the position of the rheostat. At its maximum setting, the rheostat will impose a 4V voltage drop and will therefore dissipate 1.6 watts from the full length of its resistive element. If the rheostat is reset to impose only a 4-ohm resistance, the voltage drop that it imposes will be 2V, the current in the circuit will be 0.5 amps, and the rheostat will therefore dissipate 1 watt from 4/10ths of the length of its resistive element. A

wire-wound potentiometer will be better able to handle high dissipation from a short segment of its element than other types of rheostat. Add a fixed resistor in series with a rheostat if necessary to impose a limit on the current.

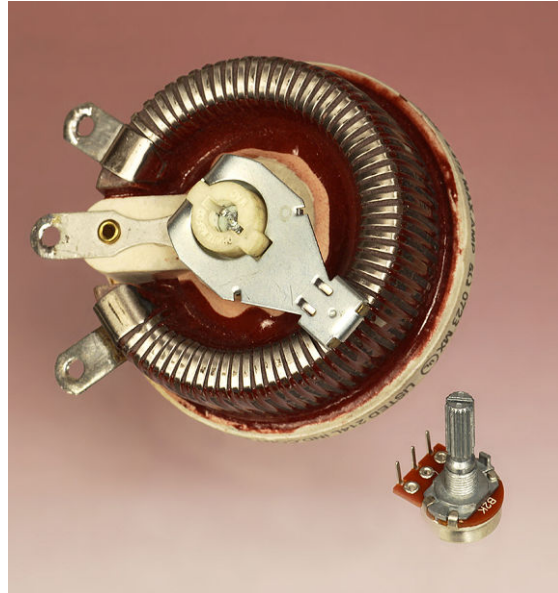


Figure 11-13. The large potentiometer is approximately 3" in diameter, rated at 5 ohms, and able to handle more than 4 amps. The small potentiometer is 5/8" diameter, rated at 2K and 1/4 watt, with pins designed for through-hole insertion in a circuit board, and a grooved shaft that accepts a push-on knob. Despite the disparity in size, the principle of operation and the basic features are identical.

When using a trimmer potentiometer, limit the current through the wiper to 100mA as an absolute maximum value.

The Wrong Taper

When buying a potentiometer, remember to check the specification to find out whether it has linear or audio/log taper. If necessary, attach a meter, with the potentiometer set to its center position, to verify which kind of taper you have. While holding the meter probes in place, rotate the potentiometer shaft to determine which way an audio/log taper is oriented.

capacitor

12

Quite often referred to as a *cap*. Formerly known (primarily in the United Kingdom) as a *condenser*, but that term has become obsolete.

OTHER RELATED COMPONENTS

- **variable capacitor** (See [Chapter 13](#))
- **battery** (See [Chapter 2](#))

What It Does

A capacitor connected across a DC power source will accumulate a charge, which then persists after the source is disconnected. In this way, the capacitor stores (and can then discharge) energy like a small rechargeable battery. The charge/discharge rate is extremely fast but can be limited by a series resistor, which enables the capacitor to be used as a timing component in many electronic circuits.

A capacitor can also be used to block DC current while it passes pulses, or electrical “noise,” or alternating current, or audio signals, or other wave forms. This capability enables it to smooth the output voltage provided by **power supplies**; to remove spikes from signals that would otherwise tend to cause spurious triggering of components in digital circuits; to adjust the frequency response of an audio circuit; or to couple separate components or circuit elements that must be protected from transmission of DC current.

Schematic symbols for capacitors are shown in [Figure 12-1](#). At top-left is a nonpolarized capacitor, while the other two indicate that a polarized capacitor must be used, and must be oriented as shown. The variant at the bottom is most commonly used in Europe. Confusingly, the nonpolarized symbol may also be used to identify a po-

larized capacitor, if a + sign is added. The polarized symbols are sometimes printed without + signs, but the symbols still indicate that polarity must be observed.

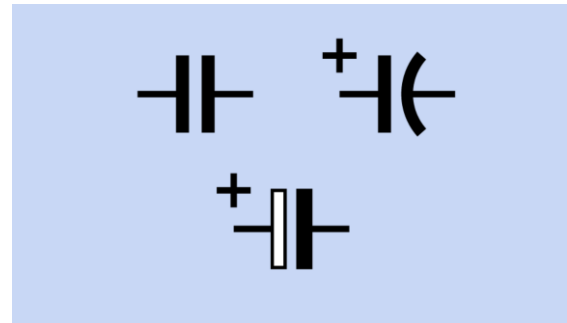


Figure 12-1. Schematic symbols for polarized and nonpolarized capacitors. See text for details.

How It Works

In its simplest form, a capacitor consists of two plates, each with a lead attached to it for connection with a DC power source. The plates are separated by a thin, insulating layer known as the *dielectric*, which is usually a solid or a paste but may be liquid, gel, gaseous, or vacuum.

The plates in most capacitors are made from thin metal film or metallized plastic film. To minimize the size of the component, the film may be rolled up to form a compact cylindrical package, or multiple flat sections may be interleaved.

Electrons from the power source will migrate on to the plate attached to the negative side of the source, and will tend to repel electrons from the other plate. This may be thought of as creating *electron holes* in the other plate or as attracting *positive charges*, as shown in Figure 12-2. When the capacitor is disconnected from the power source, the opposite charges on its plates will persist in equilibrium as a result of their mutual attraction, although the voltage will gradually dissipate as a result of *leakage*, either through the dielectric or via other pathways.

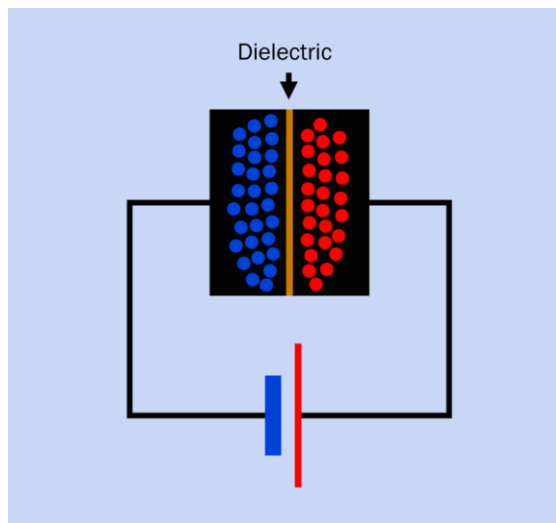


Figure 12-2. Because the plates of a capacitor are electrically conductive, they will become populated with positive and negative charges when connected with a DC power source. As opposite charges attract each other, they will tend to congregate on either side of the *dielectric*, which is an insulating layer. The battery symbol is shown here colored for clarity.

When a **resistor** is placed across the leads of a charged capacitor, the capacitor will discharge

itself through the resistor at a rate limited by the resistance. Conversely, if a capacitor is charged through a resistor, the resistor will limit the charging rate.

A resistor in series with a capacitor is known as an *RC network* (Resistor-Capacitor network). In Figure 12-3, an RC circuit is shown with a SPDT switch that charges or discharges the capacitor via a series resistor. The voltage at point A increases nonlinearly (relative to the negative side of the power supply) while the capacitor is charging, and decreases nonlinearly while the capacitor is discharging, as suggested by the graphs. At any moment, the time that the capacitor takes to acquire 63% of the difference between its current charge and the voltage being supplied to it is known as the *time constant* for the circuit. See “The Time Constant” (page 105) for additional information.

When a capacitor is connected across an AC voltage source, each surge of electrons to one plate induces an equal and opposite positive surge to the other plate, and when polarity of the power supply reverses, the charges on the plates switch places. These surges may make it seem that the capacitor is conducting AC current, even though the dielectric that separates the plates is an insulator. See Figure 12-4. Often a capacitor is said to “pass” AC, even though this is not really happening. For convenience, and because the concept is widely established, this encyclopedia refers to capacitors as “passing” AC.

Depending on the size of the capacitor, it will block some AC frequencies while passing others. Generally speaking, a smaller capacitor will pass high frequencies relatively efficiently, as each little surge of current fills each plate. However, the situation is complicated by the *inductive reactance* (which creates the *effective series resistance*) of a capacitor, as discussed below. See “Alternating Current and Capacitive Reactance” (page 106).

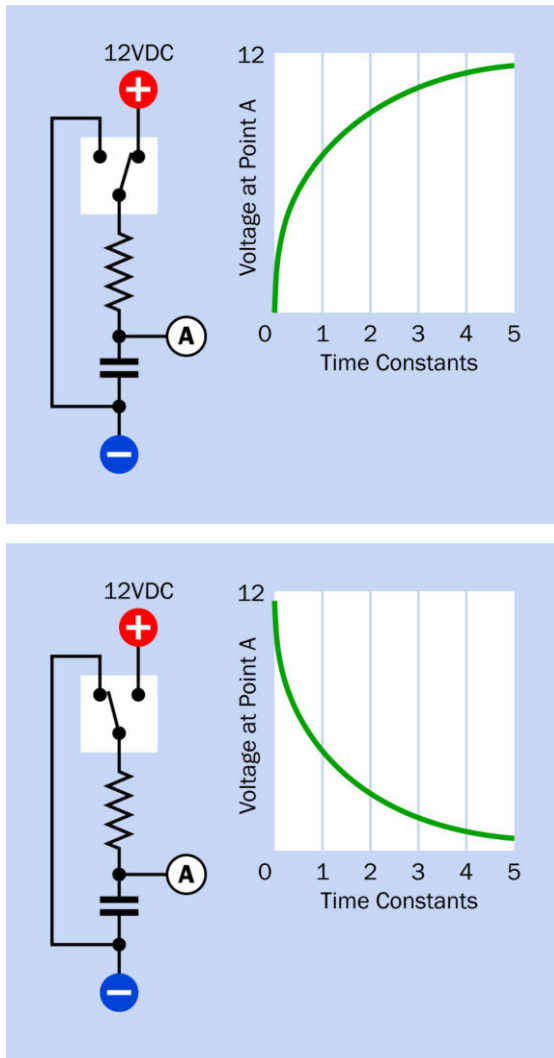


Figure 12-3. An RC (Resistor-Capacitor) network with a switch to control charge and discharge of a capacitor. At top, the curve gives an approximate idea of the charging behavior of the capacitor. At bottom, the curve illustrates its discharging behavior.

Variants

Format

The three most common packages for capacitors are cylindrical, disc, and rectangular tablet.

A *cylindrical capacitor* may have *axial leads* (a wire attached to each end) or *radial leads* (both wires emerging from one end). Radial capacitors are

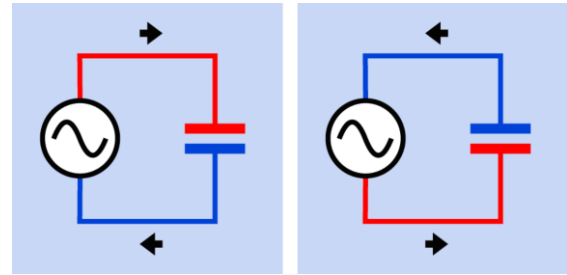


Figure 12-4. In the left diagram, a source of alternating current charges the upper plate of a capacitor positively and the lower plate negatively. This process entails a flow of conventional current shown by the arrows. A moment later, when the AC current flow reverses, the flow also reverses, creating the impression that the capacitor “passes” AC current.

more widely used as they allow easy insertion into a circuit board. The capacitor is usually packaged in a small aluminum can, closed at one end, capped with an insulating disc at the other end, and wrapped in a thin layer of insulating plastic. Some samples are shown in [Figure 12-5](#) and [Figure 12-6](#).

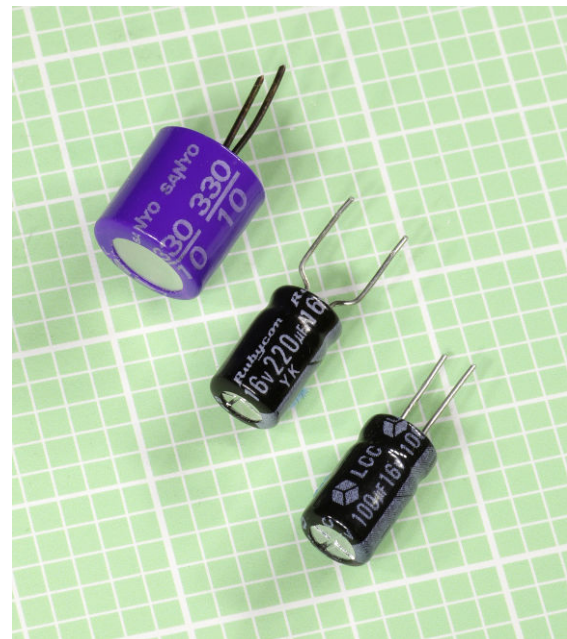


Figure 12-5. Cylindrical capacitors with radial leads. All are electrolytic.

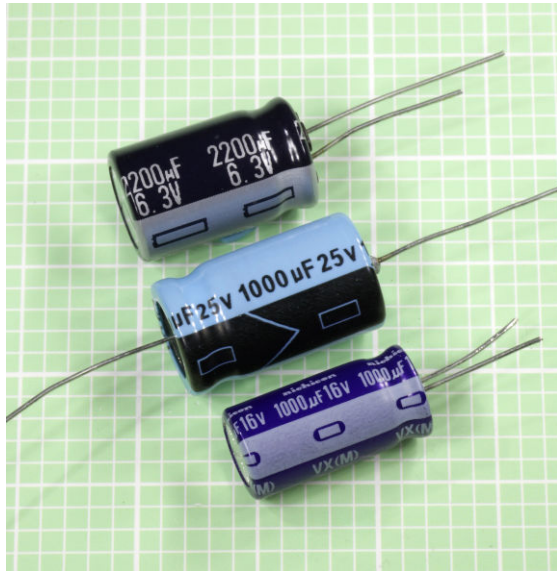


Figure 12-6. Cylindrical capacitors with radial leads (top and bottom) and axial leads (center). All are electrolytic.

A *disc capacitor* (sometimes referred to as a *button capacitor*) is usually encased in an insulating ceramic compound, and has radial leads. Modern small-value ceramic capacitors are more likely to be dipped in epoxy, or to be square tablets. Some samples are shown in [Figure 12-7](#).

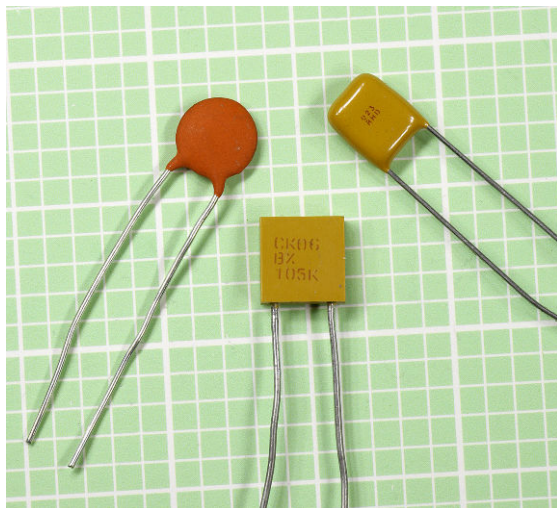


Figure 12-7. Generic ceramic capacitors. Left: rated for 0.1µF at 50V. Center: 1µF at 50V. Right: 1µF at 50V.

A *surface-mount capacitor* is square or rectangular, usually a few millimeters in each dimension, with two conductive pads or contacts at opposite ends. It appears almost identical to a surface-mount resistor. Larger-value capacitors are inevitably bigger but can still be designed for surface-mount applications. See [Figure 12-8](#).



Figure 12-8. Most surface-mount capacitors are as tiny as other surface-mount components, but this 4,700µF electrolytic (at 10V) has a base approximately 0.6" square. A solder tab is visible at the center of the nearest edge.

Many capacitors are nonpolarized, meaning that they are insensitive to polarity. However, electrolytic and tantalum capacitors must be connected “the right way around” to any DC voltage source. If one lead is longer than the other, it must be the “more positive” lead. A mark or band at one end of the capacitor indicates the “more negative” end. Tantalum capacitors are likely to indicate the positive lead by using a + sign on the body of the component.

An arrow printed on the side of a capacitor usually points to the “more negative” terminal. In an aluminum can with axial leads, the lead at one end will have an insulating disc around it while the other lead will be integral with the rounded end of the can. The wire at the insulated end must be “more positive” than the wire at the other end.

A **capacitor array** contains two or more capacitors that are isolated from each other internally and accessed by external contacts. They are sold in surface-mount format and also in **through-hole** chips of DIP (dual-inline package) or SIP (single-inline package) format. The internal components may be connected in one of three configurations: isolated, common-bus, or dual-ended common bus. Technically the isolated configuration should be referred to as a capacitor array, but in practice, all three configurations are usually referred to as **capacitor networks**. See [Figure 12-9](#) and [Figure 12-10](#).

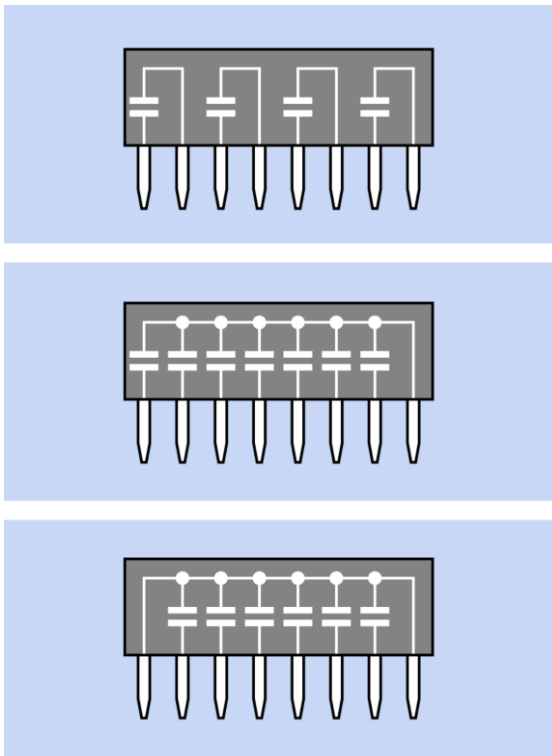


Figure 12-9. A **capacitor network** most often consists of a single-inline package (SIP) chip containing multiple capacitors in one of three configurations shown here. Top: Isolated. Center: Common bus. Bottom: Dual-ended common bus. Individual capacitor values ranging from 0.001 μ F to 0.1 μ F are common.

Capacitor networks can reduce the component count in circuits where digital logic chips require bypass capacitors. They are comparable in concept to **resistor arrays**.

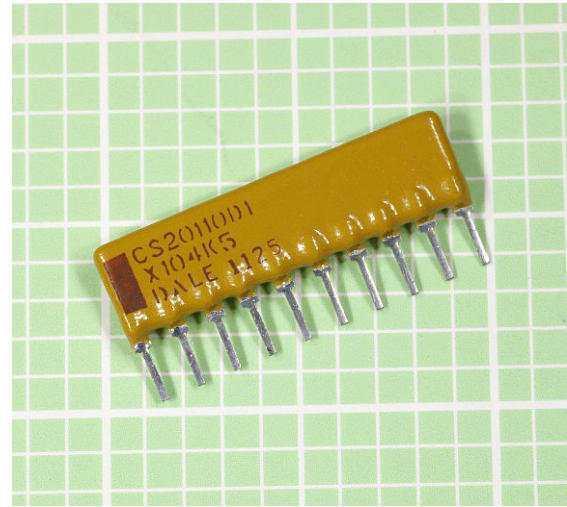


Figure 12-10. A capacitor array in through-hole, SIP format.

Chips containing **RC circuits** (multiple resistor-capacitor pairs) are available, although uncommon.

Principal Types

Electrolytic capacitors are relatively cheap, compact, and available in large values. These attributes have made them a popular choice in consumer electronics, especially for power supplies. The capacitive capability of an electrolytic is refreshed by periodic application of voltage. A moist paste inside the capacitor is intended to improve the dielectric performance when voltage is applied, but can dry out during a period of years. If an electrolytic is stored for 10 years or so, it may allow a short circuit between its leads when power is applied to it. The capacitors in [Figure 12-5](#) and [Figure 12-6](#) are all electrolytic. The capacitor in [Figure 12-11](#) is at the high end of the scale.

A **bipolar electrolytic** is a single package containing two electrolytic capacitors in series, end-to-end, with opposed polarities, so that the combination can be used where the voltage of a signal fluctuates above and below 0VDC. See [Figure 12-12](#) and [Figure 12-13](#). This type of component is likely to have “BP” (bipolar) or “NP”



Figure 12-11. This 13,000 μ F electrolytic capacitor is larger than would be required in most everyday applications.

(nonpolarized) printed on its shell. It may be used in audio circuits where polarized capacitors are normally unsuitable, and is likely to be cheaper than non-electrolytic alternatives. However, it suffers from the same weaknesses as all electrolytics.

Tantalum capacitors are compact but relatively expensive, and can be vulnerable to voltage spikes. They are sensitive to application of the wrong polarity. Typically they are epoxy-dipped rather than mounted inside a small aluminum can like electrolytics, and consequently the electrolyte may be less likely to evaporate and dry out. In [Figure 12-14](#), two tantalum capacitors (rated 330 μ F at 6.3V, left, and 100 μ F at 20V, right) are shown above a polyester film capacitor (rated

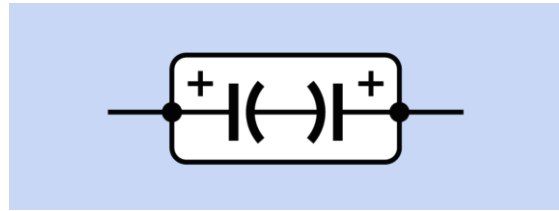


Figure 12-12. Schematic view of the internal configuration of a bipolar electrolytic capacitor, also known as non-polarized electrolytic capacitor. It consists of two electrolytics in series, with opposing polarities.

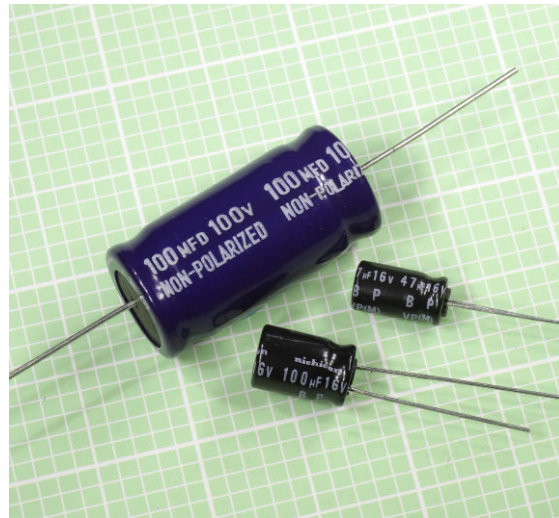


Figure 12-13. Bipolar electrolytic capacitors. The larger size of the one at top-left is a consequence of its higher voltage rating. “BP” on the other two capacitors is an acronym for “bipolar,” meaning that they have no polarity, even though one lead may be shorter than the other.

10 μ F at 100V). Surface-mount tantalum capacitors are decreasing in popularity as large-value ceramic capacitors are becoming available, with smaller dimensions and lower equivalent series resistance.

Plastic-film capacitors are discussed in the following section.

Single-layer **ceramic** capacitors are often used for bypass, and are suitable for high-frequency or audio applications. Their value is not very stable with temperature, although the “NPO” variants are more stable. Multilayer ceramic capacitors are more compact than single-layer ceramic, and

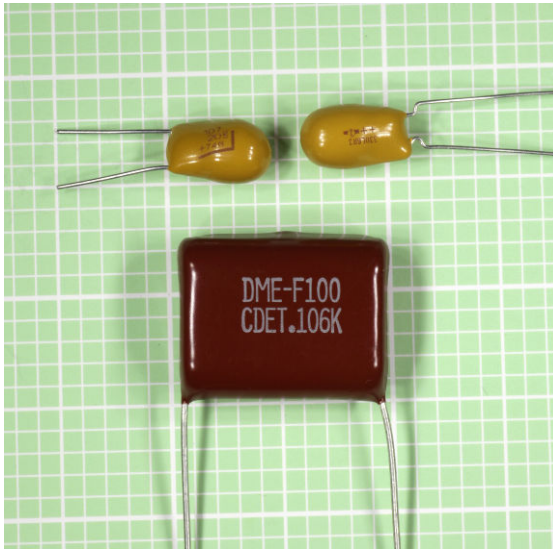


Figure 12-14. Two tantalum capacitors are shown above a polyester film capacitor. The polarity of the tantalum capacitors is indicated by the plus signs adjacent to the longer lead, in each case. The polyester capacitor is non-polarized.

consequently are becoming increasingly popular. Three multilayer ceramic capacitors are shown in Figure 12-15. At bottom-right, even the largest (rated at $47\mu\text{F}$ at 16V) is only 0.2" square.

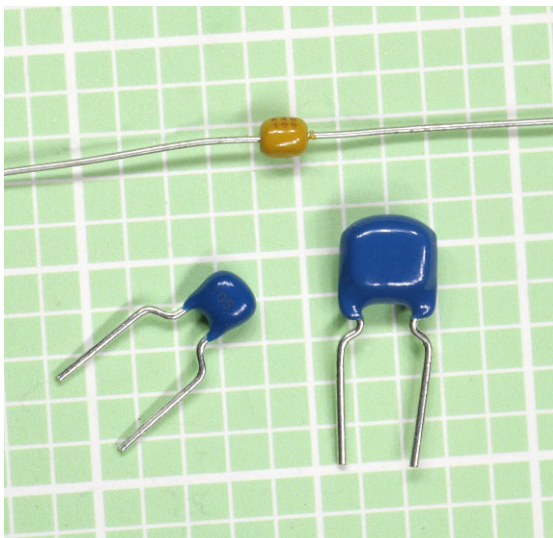


Figure 12-15. Multilayer ceramic capacitors are extremely compact, and are nonpolarized. Top: $1,000\text{pF}$ (i.e. 1nF) at 100V. Bottom left: $1\mu\text{F}$ at 25V. Bottom right: $47\mu\text{F}$ at 16V.

Dielectrics

The dielectric used in a capacitor most often consists of an electrolytic layer, a ceramic compound, a plastic film (polycarbonate, polypropylene, or polystyrene), or paper.

An *electrolytic layer* in an electrolytic capacitor traditionally consists of paper soaked in an electrolyte. It is interleaved with a thin film of aluminum on which is deposited a layer of aluminum oxide. The layers are rolled up to create a cylindrical component. The functioning dielectric is created when voltage is applied.

Polyester

This is the most common type of plastic film, with the highest dielectric constant, enabling highest capacitance per unit volume. Widely used in DC applications, but the rolled layers create parasitic inductance. Often used in decoupling, coupling, and bypass, but not so suitable for situations requiring stability and low leakage. May not be suitable for high current.

Polycarbonate

Thermally very stable, often specified for filters and timing circuits that require a fixed frequency. An excellent type of capacitor, compatible for mil-spec applications, but expensive.

Mylar, Polyester, and other plastic-film types are often used in audio circuits, where their voltage limitation (typically less than 100VDC) is not a problem, and their nonpolarized attribute is an advantage.

Polypropylene

Vulnerable to heat (a maximum of 85 degrees Centigrade is common), and less thermally stable than polycarbonate. A very low power dissipation factor allows it to handle higher power at higher frequencies. Available with tolerances down to 1%. These capacitors are a popular choice in crossover

networks for loudspeaker combinations, and are used in switching power supplies. They tend to be physically larger than other capacitors using film dielectric.

Values

Farads

The electrical storage capacity of a capacitor is measured in *farads*, universally represented by the letter F. A capacitor that can be charged with a potential difference between its plates of 1 volt, in a time of 1 second, during which it draws 1 amp, has a capacitance of 1 farad.

Because the farad is a large unit, capacitors in electronic circuits almost always have fractional values: *microfarads* (μF), *nanofarads* (nF), and *picofarads* (pF). The Greek letter μ (mu) should be used in the μF abbreviation, but a lowercase letter u is often substituted. Thus, for example, 10uF means the same as 10 μF .

1F = 1,000,000 μF , and 1 μF = 1,000,000pF. Therefore, 1 farad is equivalent to 1 trillion picofarads—a very wide range of possible values. See [Figure 12-16](#) and [Figure 12-17](#) for charts showing equivalent values in different units.

pF	nF	μF
1	0.001	0.000001
10	0.01	0.00001
100	0.1	0.0001
1,000	1	0.001
10,000	10	0.01
100,000	100	0.1
1,000,000	1,000	1

Figure 12-16. Equivalent values for picofarads, nanofarads, and microfarads. The nF unit is used primarily in Europe.

μF	F
1	0.000001
10	0.00001
100	0.0001
1,000	0.001
10,000	0.01
100,000	0.1
1,000,000	1

Figure 12-17. Equivalent values for microfarads and farads. Because the farad is such a large unit, electronic circuits almost always use fractional values.

The nF unit is more common in Europe than in the United States. A 1nF capacitance is often expressed in the US as 0.001 μF or 1,000pF. Similarly, a 10nF capacitance is almost always expressed as a 0.01 μF , and a 0.1nF capacitance is more likely to be expressed as 100pF.

European schematics may use value-symbols as a substitute for decimal points. For example, a 4.7pF capacitor may be shown as 4p7, a 6.8nF capacitor may be shown as 6n8, and a 3.3 μF capacitor may be shown as 3 μ 3.

Commonly Used Values

The traditional range of capacitor values was established on the same basis as the traditional range of resistor values, by assuming an accuracy of plus-or-minus 20% and choosing factors that would minimize the possible overlap between adjacent tolerance ranges. The factors 1.0, 1.5, 2.2, 3.3, 4.7, 6.8, and 10 satisfy this requirement. See [Chapter 10](#) for a more detailed explanation, including a graphical representation of values and overlaps in [Figure 10-8](#). While many resistors are now manufactured with high precision, 20%

tolerance is still common for electrolytic capacitors. Other types of capacitors are available with an accuracy of 10% or 5%, but are more expensive.

While large-value capacitors are likely to have their actual value printed on them, smaller capacitors are identified by a variety of different codes. These codes are not standardized among manufacturers, and exist in various colors and abbreviations. A multimeter that can measure capacitance is a quicker, easier, and more reliable method of determining the value of a component than trying to interpret the codes.

In addition to capacitance, a large capacitor is likely to have its working voltage printed on it. Exceeding this value increases the risk of damaging the dielectric. In the case of electrolytic capacitors, a voltage that is much lower than the rated value should also be avoided, because these capacitors require an electrical potential to maintain their performance.

In common electronics applications, values larger than 4,700 μ F or smaller than 10pF are unusual.

Electrolytics are available at a moderate price in a wider range of values than other commonly used capacitors. They range from 1 μ F to 4,700 μ F and sometimes beyond. Working voltages typically range from 6.3VDC to 100VDC, but can be as high as 450VDC.

Tantalum capacitors are usually unavailable in sizes above 150 μ F or for voltages above 35VDC.

Single-layer ceramic capacitors have small values ranging from 0.01 μ F to 0.22 μ F, with working voltages usually not exceeding 50VDC, although very small-value capacitors may be rated much higher for special applications. Poor tolerances of +80% to -20% are common.

Some variants of multi-layer ceramic capacitors are capable of storing up to 47 μ F, although 10 μ F is a more common upper limit. They are seldom rated above 100VDC. Some are accurate to plus-or-minus 5%.

Dielectric Constant

If A is the area of each plate in a capacitor (measured in square centimeters), and T is the thickness of the dielectric (measured in centimeters), and K is the *dielectric constant* of the capacitor, the capacitance, C (measured in farads) will be obtained from the formula:

$$C = (0.0885 * K * A) / T$$

The dielectric constant of air is 1. Other dielectrics have different standard values. Polyethylene, for instance, has a constant of approximately 2.3. Thus a capacitor of 1 square centimeter plate area and polyethylene dielectric 0.01 centimeters thick would have a capacitance of about 20pF. A tantalum capacitor of equal plate area and dielectric thickness would have capacitance closer to 100pF, since the dielectric constant of tantalum oxide is much higher than that of polyethylene.

The Time Constant

When a capacitor is charged in series through a resistor (it is used in an RC network), and it begins with no charge on its plates, the *time constant* is the time, in seconds, required to charge the capacitor to 63% of the supply voltage. After an additional, identical interval of time, the capacitor will acquire 63% of the remaining difference between itself and the power supply. In theory the capacitor gets closer and closer to a full charge, but never quite reaches 100%. However, five time constants are sufficient for the capacitor to reach 99%, which is regarded as close enough to being fully charged for all practical purposes.

Refer to [Figure 12-3](#) for a schematic of an RC network.

The time constant is a simple function of the resistance and the capacitance. If R is the value of the resistor (in ohms), and C is the value of the capacitor (in farads), the time constant, TC, will be obtained by the formula:

$$TC = R * C$$

If we multiply the R value by 1,000 while dividing the C value by 1,000, the time constant remains the same, and we can use the more convenient values of kilohms for the resistance and μF for the capacitance. In other words, the formula tells us that a 1K resistor in series with a 1,000 μF capacitor has a time constant of 1 second.

The formula suggests that if the value of R diminishes to zero, the capacitor will charge instantly. In reality, the charging time will be rapid but finite, limited by factors such as the electrical resistance of the materials used.

Multiple Capacitors

When two or more capacitors are wired in parallel, their total capacitance is the sum of their separate capacitances. When two or more capacitors are wired in series, the relationship between their total capacitance © and their individual capacitances (C1, C2, C3 . . .) is given by this formula:

$$1 / C = (1/C1) + (1/C2) + (1/C3) . . .$$

The formula to calculate the total capacitance of capacitors connected in series resembles the one used to calculate the total resistance of resistors connected in parallel. See [Chapter 10](#).

Alternating Current and Capacitive Reactance

The apparent resistance of a capacitor to AC is properly known as *capacitive reactance*. In the following formula, capacitive reactance (X_C , in ohms) is derived as a function of capacitance (C, in farads) and AC frequency (f, measured in hertz):

$$X_C = 1 / (2 * \pi * f * C)$$

The formula shows that when frequency becomes zero, capacitive reactance becomes infinite; in other words, a capacitor has theoretically infinite resistance when DC current tries to flow through it. In reality, a dielectric has a finite resistance, and thus always allows some leakage.

The formula also shows that capacitive reactance diminishes when the size of the capacitor increa-

ses and/or the frequency being applied to it increases. From this it appears that an AC signal will be attenuated less at higher frequencies, especially if we use a small capacitor. However, a real-world capacitor also exhibits some degree of *inductive reactance*. This value will depend on its configuration (cylindrical vs. multiple flat plates), its physical length, the materials from which it is fabricated, the lengths of its leads, and other factors. Inductive reactance tends to *increase* with frequency, and since capacitive reactance tends to *decrease* with frequency, at some point the curves for the two functions intersect. This point represents the capacitor's *self-resonant frequency*, which is often referred to simply as its *resonant frequency*. See [Figure 12-18](#).

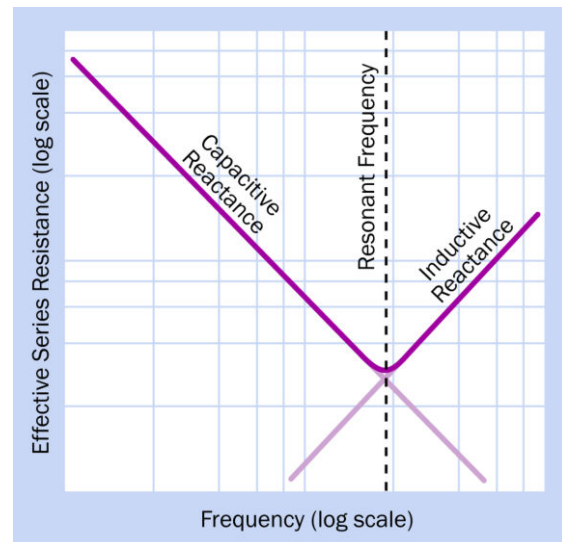


Figure 12-18. As an AC current applied to a capacitor increases in frequency, the *capacitive reactance* of the component decreases, while its *inductive reactance* increases. The resonant frequency of the capacitor is found where the two functions intersect.

Equivalent Series Resistance

A theoretically ideal capacitor would be purely reactive, without any resistance. In reality, capacitors are not ideal, and they have *equivalent series resistance*, or *ESR*. This is defined as the resistor

that you would have to place in series with an ideal version of the capacitor, so that the combination would behave like the real version of the capacitor on its own.

If X_c is the reactance of the capacitor, then its Q factor (which means its *quality factor*) is given by the simple formula:

$$Q = X_c / \text{ESR}$$

Thus, the quality factor is higher if the ESR is relatively low. However, the reactance of the capacitor will vary significantly with frequency, and this simple formula is only an approximate guide.

The Q-factor for capacitors should not be confused with the Q-factor for inductors, which is calculated quite differently.

How to Use it

The figures illustrate some simplified schematics for common applications.

Bypass Capacitor

In Figure 12-19, a low-value capacitor (often $0.1\mu\text{F}$) is placed near the power input pin of a sensitive digital chip to divert high-frequency spikes or noise to negative ground. This *bypass capacitor* may also be described as a *decoupling capacitor*.

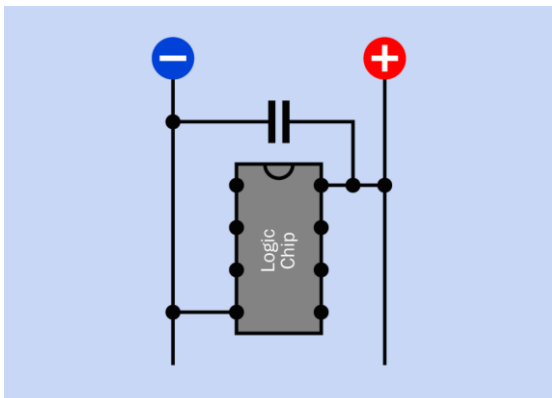


Figure 12-19. A bypass capacitor (typically $0.1\mu\text{F}$) configured to protect an integrated circuit logic chip from voltage spikes and noise in the power supply.

Coupling Capacitor

In Figure 12-20, a $1\mu\text{F}$ *coupling capacitor* transmits a pulse from one section of a circuit to another, while blocking the DC voltage. Some re-shaping of the waveform may occur.

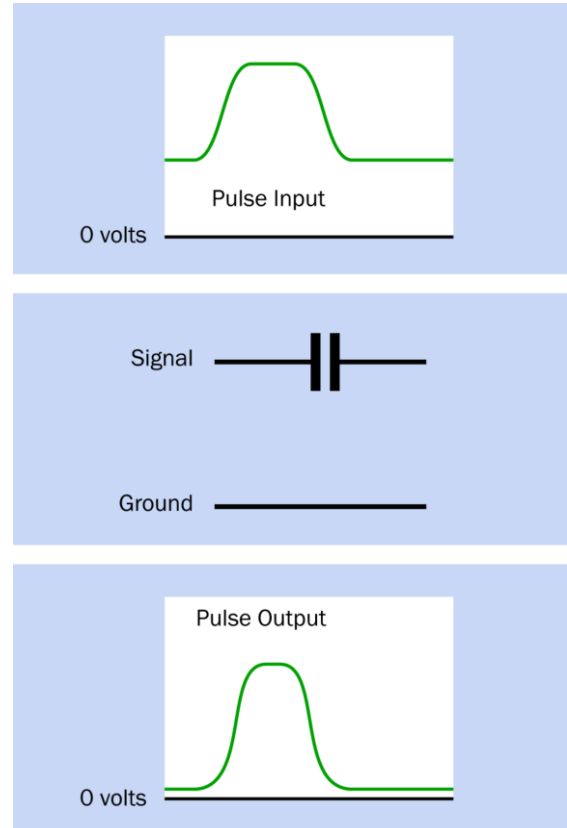


Figure 12-20. A coupling capacitor (typically around $1\mu\text{F}$) preserves DC isolation of one section of a circuit from another, while allowing a pulse to be transmitted.

High-Pass Filter

In Figure 12-21, a $0.1\mu\text{F}$ capacitor blocks the low-frequency component of a complex waveform and transmits only the higher frequency that was superimposed on the low frequency.

Low-Pass Filter

In Figure 12-22, a $0.1\mu\text{F}$ decoupling capacitor diverts the higher frequency component of a complex waveform to negative ground, allowing only

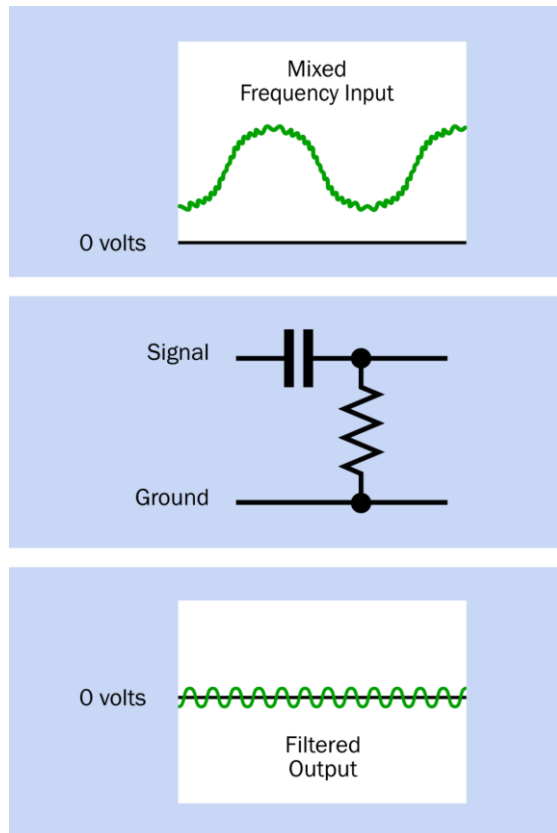


Figure 12-21. A small capacitor (typically $0.1\mu\text{F}$) can be used to create a high-pass filter, passing high frequencies while blocking low frequencies.

the lower frequency to be preserved. A lower-value capacitor (such as $0.001\mu\text{F}$) will bleed away high-frequency noise from an AM radio source without affecting audio frequencies.

Smoothing Capacitor

In [Figure 12-23](#), a $100\mu\text{F}$ capacitor charges and discharges to smooth an AC signal after a diode has removed the negative portion.

Snubber

In [Figure 12-24](#), an RC network (inside a white dashed line) is known as a *snubber* when used to protect a switch from the problem of *arcing* (pronounced “arcking”)—that is, a sustained spark that can quickly erode the switch contacts. Arcing may occur in switches, pushbuttons, or relays

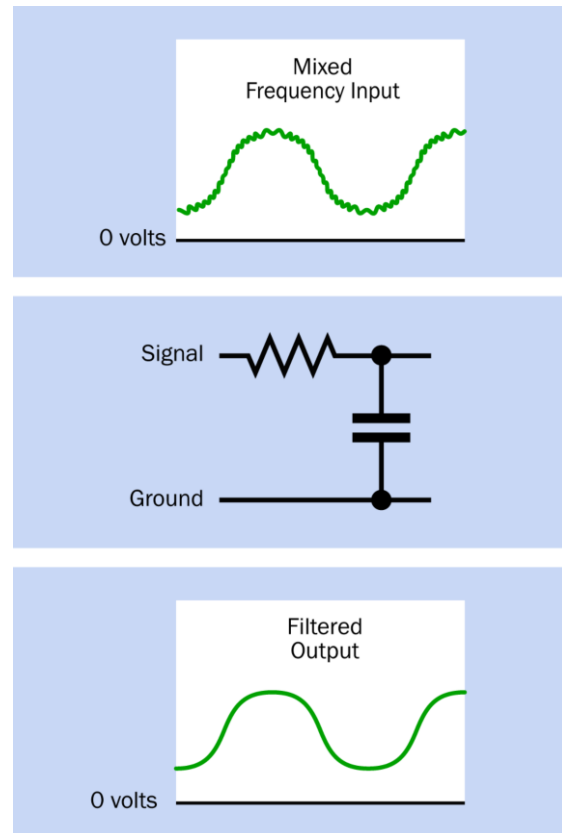


Figure 12-22. A small capacitor (typically $0.1\mu\text{F}$) in this configuration routes high frequencies to negative ground, filtering them out of an analog signal.

that control an inductive load, such as a large motor. This problem can become significant at high DC currents (10A or more) or relatively high AC or DC voltages (100V or more).

When the switch is opened, the magnetic field that has been sustained by the inductive load collapses, causing a surge of current, or *forward EMF*. The capacitor in the snubber absorbs this surge, thus protecting the switch contacts. When the switch is closed again, the capacitor discharges itself, but the resistor limits the outrush of current—again, protecting the switch.

A snubber placed around the switch in a DC circuit could typically use a $0.1\mu\text{F}$ capacitor (polypropylene or polyester) rated for 125VAC/200VDC, and a 100-ohm carbon resistor rated 0.5

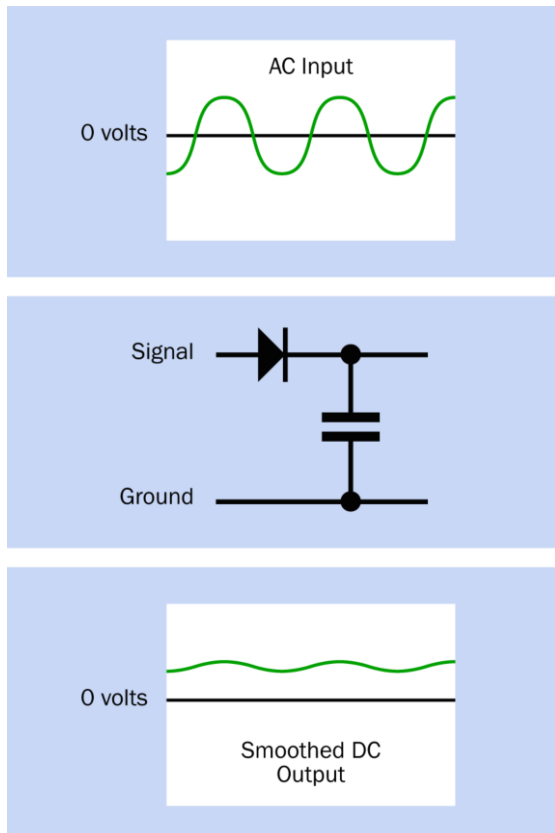


Figure 12-23. A capacitor of 100 μ F or more smooths the upper half of an AC signal that has passed through a diode. The capacitor charges during each positive pulse and discharges to “fill the gaps” between them.

watt or higher. Prepackaged snubbers containing appropriate capacitor-resistor pairs are available from some parts suppliers, primarily for industrial use.

In an AC circuit, a snubber can be placed around the inductive load itself. Although a **diode** is often used this way in a DC circuit, it cannot be used with AC.

Although solid-state switching devices such as a **solid state relay** contain no mechanical contacts, they may still be damaged by substantial pulses of back-EMF, and can be protected by a snubber where they are controlling inductive loads that take 10A or more at 100V or more.

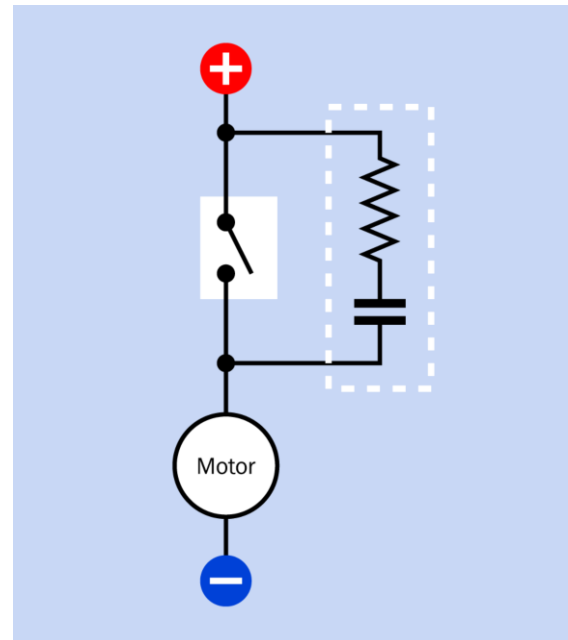


Figure 12-24. An RC network (outlined with a white dashed line) protects a switch that controls a high inductive load. Used in this way, the RC network is known as a *snubber*.

Capacitor as a Battery Substitute

A capacitor may be substituted for a battery for some applications, although it has a lower energy density and will be more expensive to manufacture. A capacitor charges and discharges much more rapidly than a battery because no chemical reactions are involved, but a battery sustains its voltage much more successfully during the discharge cycle.

Capacitors that can store a very large amount of energy are often referred to as *supercapacitors*.

What Can Go Wrong

Common problems associated with capacitors are age-related deterioration (especially in electrolytics), inductive reactance (especially in cylindrical formats), nonlinear response, resistivity, excessive current leakage, and dielectric memory. Some of these problems are discussed below. A manufacturer’s datasheet should be consulted

carefully in conjunction with the notes regarding compositions in the preceding Variants section before making a commitment to a particular type of capacitor.

Wrong Polarity

A polarized capacitor may offer virtually no resistance if it is connected the wrong way around to a DC power source. A very high current can result, damaging the capacitor and probably other components in the circuit. Failing to observe the polarity of a tantalum capacitor can have destructive or even explosive consequences, depending on the amperage.

Voltage Overload

If the DC working voltage of a capacitor is exceeded, there is a risk of breaking down the dielectric and allowing a spark, or arc, that will form a short circuit. Note that the DC rating of a capacitor does not mean that it can be used safely with an equivalent AC voltage. The maximum AC voltage should be no greater than approximately 0.7 times the DC rated voltage. If a DC-rated capacitor is used directly across an AC power line, it will create an effective short circuit.

If capacitors are connected in series or in parallel, ideally the voltage rating for each capacitor should be the same, and certainly no less than the supply voltage.

Tantalum capacitors are easily damaged by current spikes that exceed their maximum working voltage, and are unsuitable for high-frequency coupling because of their inductance.

Leakage

Charge leakage is a problem especially associated with electrolytic capacitors, which are not suitable for storing a charge over a significant interval. Polypropylene or polystyrene film capacitors are a better choice.

Dielectric Memory

Also known as *dielectric absorption*, this is a phenomenon in which a capacitor's electrolyte dis-

plays some percentage of its former voltage after the capacitor has been discharged and then disconnected from the circuit. Single-layer ceramic capacitors especially tend to suffer from this problem.

Specific Electrolytic Issues

Electrolytic capacitors have high inductive reactance, are not manufactured to close tolerances, and deteriorate significantly with age. While other components may be stockpiled and used over a period of years, this is not a sensible policy with electrolytics.

The “capacitor plague” affecting many of these capacitors manufactured from 1999 onward provided a salutary lesson regarding their potential weaknesses. Faulty composition of the dielectric allowed it to deteriorate, liberating hydrogen gas, which eventually caused the aluminum shells of the capacitors to bulge and burst. Circuit boards from major manufacturers were affected. Because the problem took two years to become apparent, literally millions of boards with faulty capacitors had been sold before the fault was diagnosed and eventually corrected.

Unfortunately electrolytics cannot be easily replaced with other types of capacitors in applications such as power supplies, because substitutes will be considerably larger and more expensive.

Heat

The equivalent series resistance (ESR) of a large capacitor inevitably means that it must dissipate some power as heat during use. Ripple current can also create heat. Capacitor performance will change as the temperature increases. A common maximum component temperature for electrolytic capacitors is 85 degrees Centigrade.

Vibration

In a high-vibration environment, electrolytics should be protected by clamping them mechanically in place, using a *capacitor clamp*, also known as a *c-clamp*.

Misleading Nomenclature

Rarely, in the United States, the term “mF” may be used as a probable alternative to μF . This can be a source of confusion and risk because mF is properly (but very rarely) used to mean “millifarads.” The term should always be avoided.

variable capacitor

Formerly known (primarily in the United Kingdom) as a *variable condenser*. The term is now obsolete.

OTHER RELATED COMPONENTS

- **capacitor** (See [Chapter 12](#))

What It Does

A variable capacitor allows adjustment of capacitance in much the same way that a **potentiometer** allows adjustment of resistance.

Large variable capacitors were developed primarily to tune radio receivers, in which they were known as *tuning capacitors*. Cheaper, simpler, and more reliable substitutes gradually displaced them, beginning in the 1970s. Today, they are still used in semiconductor fabrication, in RF plastic welding equipment, in surgical and dental tools, and in ham radio equipment.

Small *trimmer capacitors* are widely available and are mostly used to adjust high-frequency circuits. Many of them look almost indistinguishable from *trimmer potentiometers*.

The schematic symbols commonly used to represent a variable capacitor and a trimmer capacitor are shown in [Figure 13-1](#).

A *varactor* is a form of diode with variable capacitance, controlled by reverse voltage. See “[Varactor Diode](#)” (page 225) for this component.

How It Works

The traditional form of variable capacitor consists of two rigid semicircular plates separated by an air gap of 1mm to 2mm. To create more ca-

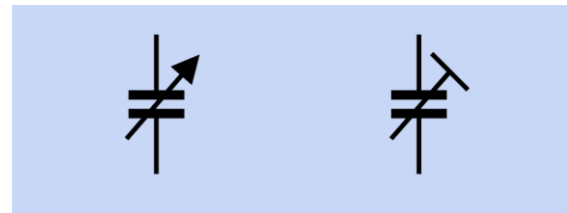


Figure 13-1. Typical schematic symbols for variable capacitor (left) and trimmer capacitor (right).

pacitance, additional interleaved plates are added to form a stack. One set of plates is known as the *rotor*, and is mounted on a shaft that can be turned, usually by an externally accessible knob. The other set of plates, known as the *stator*, is mounted on the frame of the unit with ceramic insulators. When the sets of plates completely overlap, the capacitance between them is maximized. As the rotor is turned, the sets of plates gradually disengage, and the capacitance diminishes to near zero. See [Figure 13-2](#).

The air gaps between the sets of plates are the *dielectric*. Air has a dielectric constant of approximately 1, which does not vary significantly with temperature.

The most common shape of plate is a semicircle, which provides a linear relationship between capacitance and the angle of rotation. Other shapes have been used to create a nonlinear response.

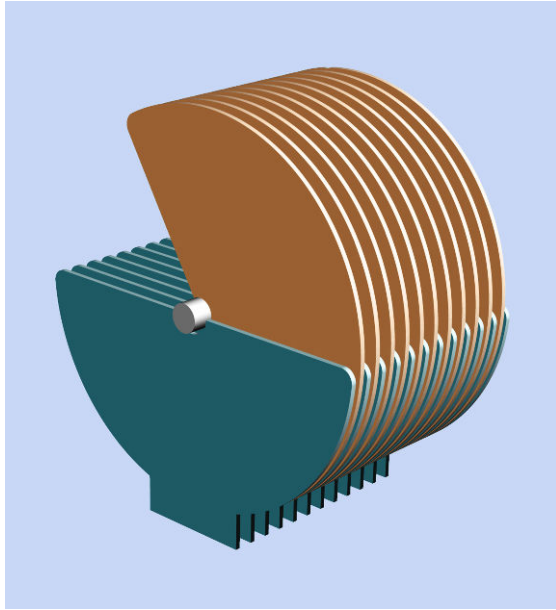


Figure 13-2. In this simplified view of a variable capacitor, the brown plates constitute the rotor, attached to a central shaft, while the blue plates are the stator. The colors have no electrical significance and are added merely for clarity. The area of overlap between rotor and stator determines the capacitance.

Reduction gears may be used to enable fine tuning of a variable capacitor, which means multiple turns of a knob can produce very small adjustments of the capacitor. At the peak of variable capacitor design, units were manufactured with high mechanical precision and included *anti-backlash gears*. These consisted of a pair of equal-sized gears mounted flat against each other with a spring between them that attempted to turn the gears in opposite directions from each other. The pair of gears meshed with a single pinion, eliminating the looseness, or backlash, that normally exists when gear teeth interlock. A vintage capacitor with a spring creating anti-backlash gearing (circled) is shown in Figure 13-3. This is a two-gang capacitor—it is divided into two sections, one rated 0 to 35pF, the other rated 0 to 160pF.

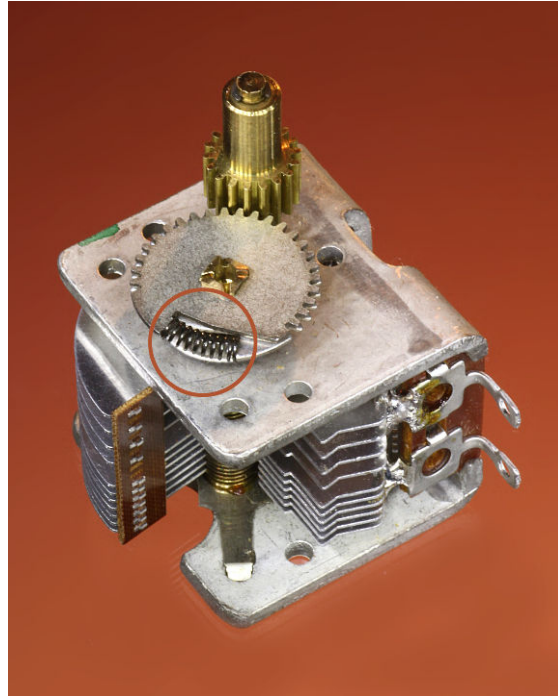


Figure 13-3. A “traditional style” variable capacitor of the type designed to tune radio frequencies. The spring, circled, enables anti-backlash gearing.

Variants

The traditional variable capacitor, with exposed, air-spaced, rigid, rotating vanes, is becoming hard to find. Small, modern variable capacitors are entirely enclosed, and their plates, or vanes, are not visible. Some capacitors use a pair of concentric cylinders instead of plates or vanes, with an external thumb screw that moves one cylinder up or down to adjust its overlap with the other. The overlap determines the capacitance.

Trimmer capacitors are available with a variety of dielectrics such as mica, thin slices of ceramic, or plastic.

Values

A large traditional capacitor can be adjusted down to a near-zero value; its maximum will be no greater than 500pF, limited by mechanical factors. (See [Chapter 12](#) for an explanation of capacitance units.)

A maximum value for a trimmer capacitor is seldom greater than 150pF. Trimmers may have their values printed on them or may be color-coded, but there is no universal set of codes. Brown, for example, may indicate either a maximum value around 2pF or 40pF, depending on the manufacturer. Check datasheets for details.

The upper limit of a trimmer's rated capacitance is usually no less than the rated value, but can often be 50% higher.

Formats

All trimmer capacitors are designed for mounting on circuit boards. Many are surface-mount, with a minority being through-hole. Surface-mount units may be 4mm × 4mm or smaller. Through-hole are typically 5mm × 5mm or larger. Superficially, trimmer capacitors resemble single-turn trimmer potentiometers with a screw head in the center of a square package. A through-hole example is shown in [Figure 13-4](#).

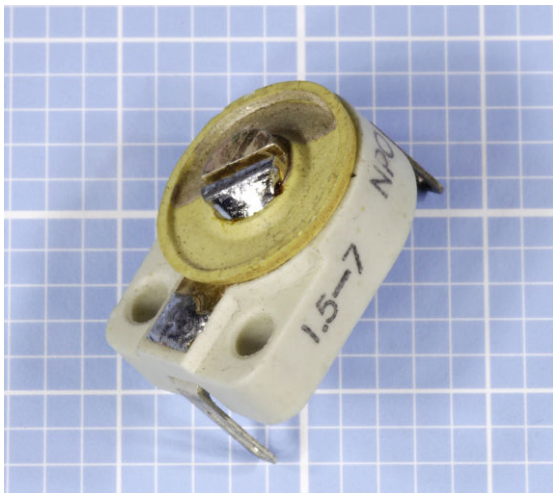


Figure 13-4. A trimmer capacitor rated 1.5pF to 7.0pF.

How to Use it

A variable capacitor is often used to tune an *LC circuit*, so called because a coil (with reactance customarily represented by letter L) is wired in parallel with a variable capacitor (represented by letter C). The schematic in [Figure 13-5](#) shows an imaginary circuit to illustrate the principle. When the switch is flipped upward, it causes a large fixed-value capacitor to be charged from a DC power source. When the switch is flipped down, the capacitor tries to pass current through the coil—but the coil's *reactance* blocks the current and converts the energy into a magnetic field. After the capacitor discharges, the magnetic field collapses and converts its energy back into electricity. This flows back to the capacitor, but with inverted polarity. The cycle now repeats with current flowing in the opposite direction. A low-current **LED** across the circuit would flash as the voltage oscillates, until the energy is exhausted.

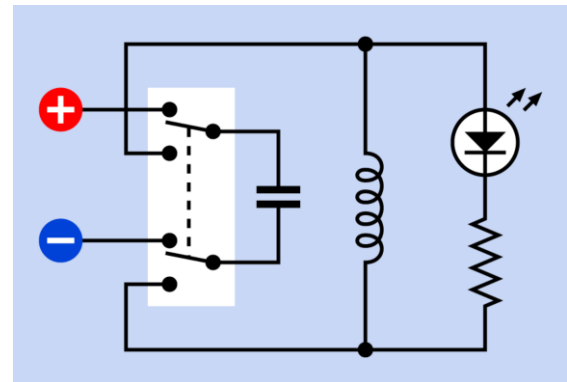


Figure 13-5. In this imaginary circuit, the capacitor is charged through the double-pole switch in its upper position. When the switch is turned, the capacitor forms an LC (inductance-capacitance) circuit with the coil, and resonates at a frequency determined by their values. In reality, extremely high values would be needed to obtain a visible result from the LED.

Because the oscillation resembles water sloshing from side to side in a tank, an LC circuit is sometimes referred to as a *tank circuit*.

In reality, unrealistically large values would be required to make the circuit function as described. This can be deduced from the following formula, where f is the frequency in Hz, L is inductance in Henrys, and C is capacitance in Farads:

$$f = 1 / (2\pi * \sqrt{L * C})$$

For a frequency of 1Hz, a massive coil opposite a very large capacitor of at least 0.1F would be needed.

However, an LC circuit is well-suited to very high frequencies (up to 1,000MHz) by using a very small coil and variable capacitor. The schematic in [Figure 13-6](#) shows a high-impedance earphone and a diode (right) substituted for the LED and the resistor in the imaginary circuit, while a variable capacitor takes the place of the fixed capacitor. With the addition of an antenna at the top and a ground wire at the bottom, this LC circuit is now capable of receiving a radio signal, using the signal itself as the source of power. The resonant frequency of the circuit is tuned by the variable capacitor. The impedance peaks at the resonant frequency, causing other frequencies to be rejected by passing them to ground. With suitable refinement and amplification, the basic principle of an LC circuit is used in AM radios and transmitters.

Because variable capacitors are so limited in size, they are unsuitable for most timing circuits.

Trimmer capacitors are typically found in high-power transmitters, cable-TV transponders, cellular base stations, and similar industrial applications.

They can be used to fine-tune the resonant frequency of an oscillator circuit, as shown in [Figure 13-7](#).

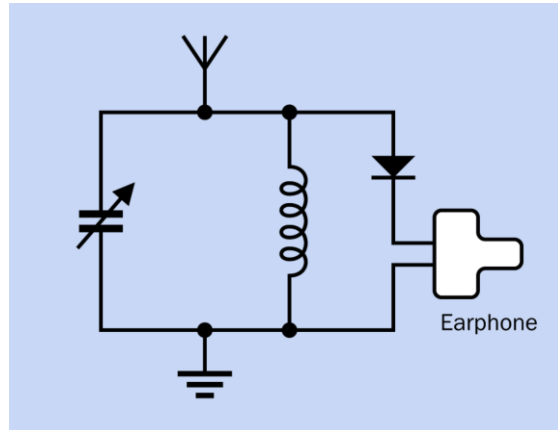


Figure 13-6. The principle of an LC circuit is used here in a basic circuit that can tune in to a radio station and create barely audible sound through the earphone at right, using only the broadcast signal for power. The variable capacitor adjusts the frequency of the circuit to resonate with the carrier wave of the radio signal.

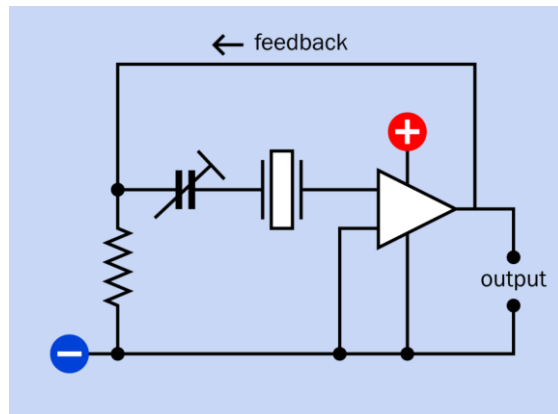


Figure 13-7. A trimmer capacitor in series with a crystal fine-tunes the frequency of this basic circuit using an op-amp.

In addition to tuning a circuit frequency, a trimmer capacitor can be used to compensate for changes in capacitance or inductance in a circuit that are caused by the relocation of wires or re-routing of traces during the development process. Readjusting a trimmer is easier than swapping fixed-value capacitors. A trimmer may also be used to compensate for capacitance in a circuit that gradually drifts with age.

What Can Go Wrong

Failure to Ground Trimmer Capacitor While Adjusting it

Although trimmer capacitors are not polarized, the manufacturer may mark one terminal with a plus sign and/or the other with a minus sign. If the capacitor is adjusted while its negative terminal is floating or ungrounded, a metal screwdriver blade will create erroneous readings. Always ground the appropriate side of a trimmer capacitor before fine-tuning it, and preferably use a plastic-bladed screwdriver.

Application of Overcoat Material or “Lock Paint”

Overcoat material is a rubbery adhesive that may be spread over assembled components to im-

munize them against moisture or vibration. *Lock paint* is a dab of paint that prevents a screw adjustment from turning after it has been set. Most manufacturers advise against applying these materials to a trimmer capacitor, because if penetration occurs, the capacitor can fail.

Lack of Shielding

Variable capacitors should be shielded during use, to protect them from external capacitive effects. Merely holding one's hand close to a variable capacitor will change its value.

inductor

The term **inductor** is used here to describe a coil that has the purpose of creating *self-inductance* in an electronic circuit, often while passing alternating current in combination with resistors and/or capacitors. A *choke* is a form of inductor. By comparison, the **electromagnet** entry in this encyclopedia describes a coil containing a center component of ferromagnetic material that does not move relative to the coil, and has the purpose of attracting or repelling other parts that respond to a magnetic field. A coil containing a center component of ferromagnetic material that moves as a result of current passing through the coil is considered to be a **solenoid** in this encyclopedia, even though that term is sometimes more broadly applied.

OTHER RELATED COMPONENTS

- **solenoid** (See [Chapter 21](#))
- **electromagnet** (See [Chapter 20](#))

What It Does

An inductor is a coil that induces a magnetic field in itself or in a [core](#) as a result of current passing through the coil. It may be used in circuits to block or reshape AC current or a range of AC frequencies, and in this role can “tune” a simple radio receiver or various types of oscillators. It can also protect sensitive equipment from destructive voltage spikes.

The schematic symbol for an inductor includes a coil that can be drawn in two basic styles, shown at the top and at the bottom of [Figure 14-1](#). The style at the bottom has become more common. In each vertical section of the diagram, the functionality of the symbols is identical.

One or two parallel lines alongside the coil indicate that it is wound around a solid core of material that can be magnetized, while one or two

dotted lines indicate that it is wound around a core containing metal particles, such as iron filings. Where no core is shown, this indicates an air core.

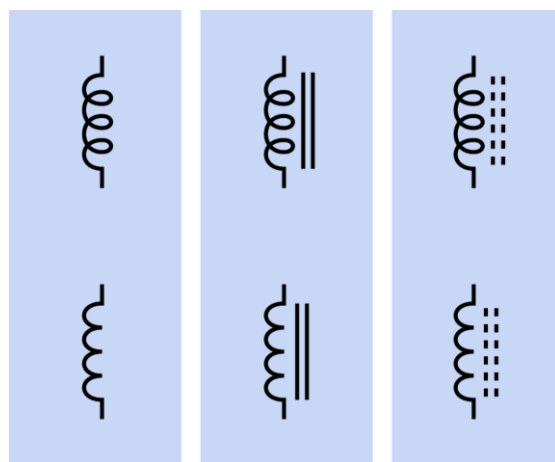


Figure 14-1. The coil symbol for an inductor may be drawn in two styles which are functionally identical. Line(s) beside the coil indicate a solid core. Dotted line(s) indicate a core containing metal particles.

A selection of inductors designed for through-hole mounting is shown in [Figure 14-2](#).

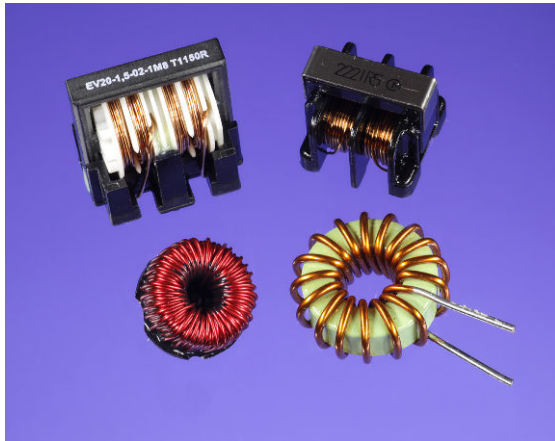


Figure 14-2. Four inductors designed for through-hole insertion into printed circuit boards.

How It Works

Direct current passing through an electrical conductor, such as a wire, creates a magnetic field around the conductor. In [Figure 14-3](#), *conventional current* (flowing from positive to negative) is passing through a straight wire from left to right, as indicated by the red/blue arrow. The resulting magnetic field is indicated by the green arrows. If the wire is now bent into a curve, as shown in [Figure 14-4](#), the magnetic field exerts an aggregate force downward through the curve. This magnetic force is conventionally said to flow from *south* to *north*.

If direct current could be induced to circulate through an unbroken circle of wire, the resulting magnetic field could exert a force through the circle as shown in [Figure 14-5](#), assuming clockwise circulation of conventional current as suggested by the red/blue arrows.

Conversely, if a magnet was pushed through the center of the circle, it would induce a pulse of electric current in the circle. Thus, electricity passing through a wire can induce a magnetic field around the wire, and conversely, a magnet moving near a wire can induce an electric current

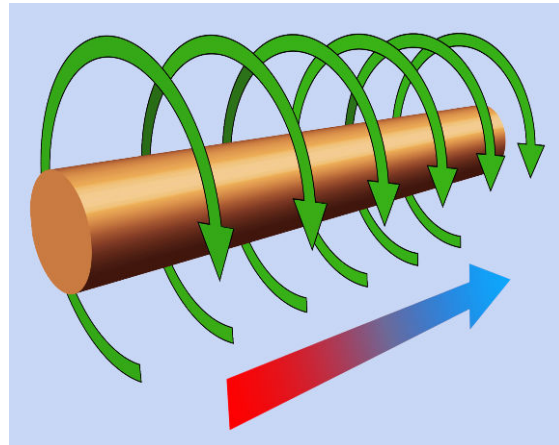


Figure 14-3. Conventional current passing through a wire from left to right (as indicated by the red/blue arrow) induces a magnetic field around the wire (shown by the green arrows).

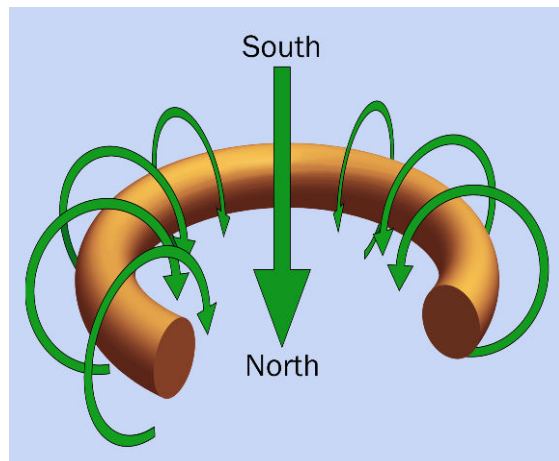


Figure 14-4. If the wire is bent into a curve, the magnetic fields can create a net force shown by the large green arrow.

in the wire. This principle is used in an electrical *generator*, and also in a **transformer**, where alternating current in the primary coil induces a fluctuating magnetic field in the core, and the field in the core is turned back into alternating current in the secondary coil.

Note that a static or unchanging magnetic field will not induce a flow of electricity.

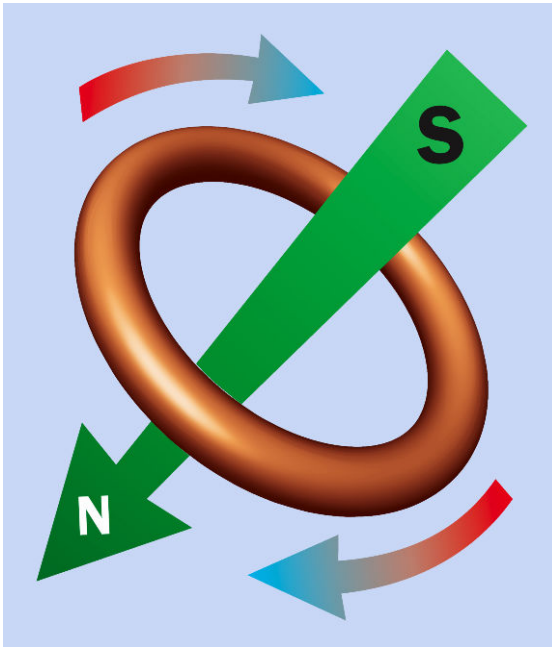


Figure 14-5. Hypothetically, if conventional current flows around a circular conductor (as suggested by the red/blue arrows), it will create a magnetic field that can create a force as shown by the green arrow.

DC Through a Coil

If the wire is formed into a helix (a series of approximate circles) as shown in Figure 14-6, and if DC current is passed through the wire, the aggregate of the magnetic fields can create a force in the direction of the green arrow in each example, depending whether the wire is wound clockwise or counter-clockwise, and depending on the direction of the current. The helix is usually referred to as a *coil* or a *winding*.

In actuality, a magnetic field is not open-ended, and its lines of force are completed by circling around outside the inductor, to complete a *magnetic circuit*. This completion of the field can be demonstrated by the traditional high-school experiment of positioning a compass or scattering iron filings on a sheet of paper above a magnet. A simplified depiction of lines of force completing a magnetic circuit is shown in Figure 14-7,

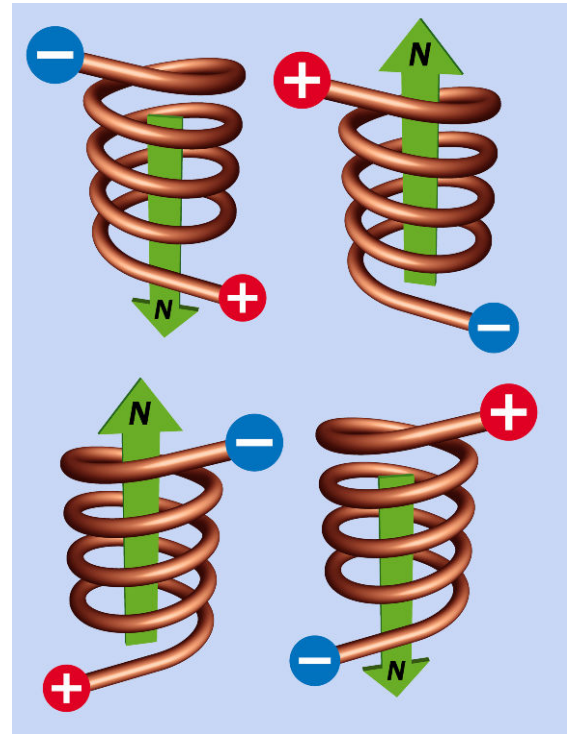


Figure 14-6. When DC current flows through a coil, it creates magnetic fields that will exert a force whose direction depends on the direction of the current and on whether the coil is wound clockwise or counterclockwise. The force is shown by the green arrow in each case.

where a coil is inducing the magnetic field. Note that throughout this encyclopedia, the color green is used to indicate the presence of magnetic force.

The completion of a magnetic field is not relevant to the primary function of the inductor. In fact the external part of the magnetic field is mostly a source of trouble in electronics applications, since it can interact with other components, and may necessitate the use of magnetic shielding. In addition, the field is weakened by completing itself through air, as air presents much greater *reluctance* (the magnetic equivalent of resistance) than the core of an electromagnet.

The polarity of a magnetic field created by a coil can be demonstrated by moving a small permanent magnet toward the coil, as shown in

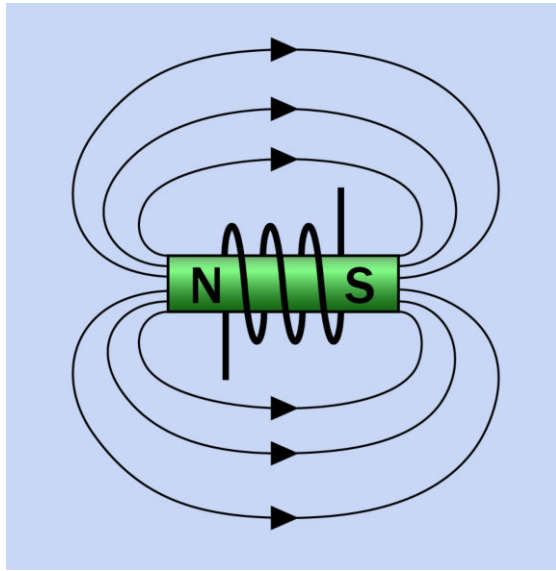


Figure 14-7. A magnetic field in reality is not open-ended, and each line of force traveling through a rod-shaped magnet or electromagnet is completed outside of the magnet. The completion of magnetic fields has been omitted from other diagrams here for clarity.

Figure 14-8. If the magnet has opposite polarity to the coil, it will tend to be repelled, as like poles repel. If it has the same polarity, it will tend to be attracted, because opposite poles attract. This principle may be used in **solenoids**.

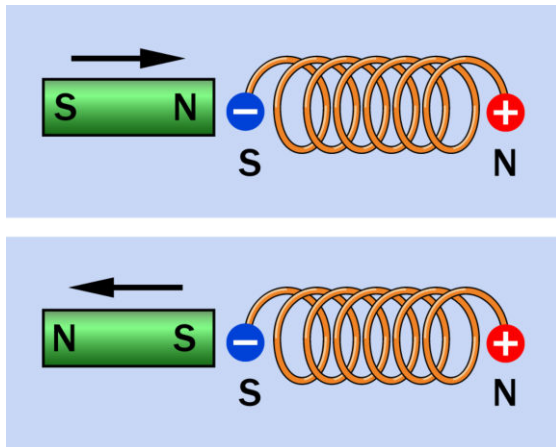


Figure 14-8. A permanent magnet (left) will either be drawn toward a DC-energized coil or repelled from it, depending on the polarity of the two magnetic fields.

Magnetic Core

The inductive power of a coil will be enhanced, and the saturation point will be reduced, by using a **magnetic core**. The term “magnetic” in this context does not mean that the core is a permanent magnet; it means that the core can be magnetized briefly by a transient pulse of electricity through the surrounding coil.

A core enhances the effectiveness of an inductor because it has a lower reluctance than that of air. In other words, magnetic flux will flow much more readily through the core than through air.

Roughly speaking, the **permeability** of a magnetic circuit is the opposite of reluctance; it is a measure of how easily a magnetic field can be induced, and is usually expressed relative to the permeability of air, which is approximately 1. The permeability of different core types is discussed in the following “Values” section.

The core of the coil contains **magnetic domains** that behave as tiny magnets, with north and south poles. In the absence of a polarizing magnetic field, the domains are randomly aligned. As a magnetic field is introduced around them and grows stronger, the domains align themselves with it, increasing the total magnetic force. When the domains are almost all uniformly aligned, the core approaches **magnetic saturation** and ceases adding to the net magnetic field. At this point the current in the inductor is said to be **continuous**.

When power to the coil is disconnected, the domains revert partially to their previous random orientation. Thus the core remains a weak permanent magnet. This effect is known as **hysteresis**, while the weak residual field is known as **remanent magnetism**.

EMF and Back-EMF

When DC current is connected through an inductor, the creation of a magnetic field takes a brief but measurable period of time. The field induces an **EMF (electro-motive force)** in the wire. Since this force opposes the supplied current, it

is referred to as *back-EMF*. It lasts only so long as the field is increasing to its full strength. After the field reaches a steady state, current flows through the coil normally.

This transient resistive effect is caused by the *self-inductance* of the coil, and is opposite to the behavior of a capacitor, which encourages an initial inrush of direct current until it is fully charged, at which point it blocks subsequent current flow.

When high-frequency alternating current attempts to flow through an inductor, if each pulse is too brief to overcome the back-EMF, the coil will block the current. A coil can thus be designed to block some frequencies but not others.

Even a simple electrical circuit that does not contain a coil will still have some self-inductance, simply because the circuit consists of wires, and even a straight length of wire induces a magnetic field when the power is switched on. However, these inductive effects are so small, they can generally be ignored in practical applications.

The transient electrical resistance to alternating current caused by either an inductor or a capacitor is known as *reactance*, although it occurs under opposite electrical conditions, as the coil impedes an initial pulse of DC current and then gradually allows it to pass, while a capacitor allows an initial pulse of DC current and then impedes it.

When a flow of DC current through a coil is switched off, the magnetic field that was created by the coil collapses and releases its stored energy. This can cause a pulse of forward EMF, and like back-EMF, it can interfere with other components in a circuit. Devices such as motors and large relays that contain substantial coils can create problematic spikes of back-EMF and forward-EMF. The forward-EMF that occurs when power to the coil is interrupted is typically dealt with by putting a diode in parallel with the coil, allowing current to circulate through it. This is known as *clamping* the voltage transient. A diode-

capacitor combination known as a *snubber* is also commonly used. For a schematic and additional information on this topic, see “*Snubber*” (page 108).

A schematic to demonstrate EMF and back-EMF is shown in Figure 14-9. The coil can be a 100-foot spool of 26-gauge (or smaller) hookup wire, or magnet wire. It will function more effectively if a piece of iron or steel, such as half-inch galvanized pipe, is inserted through its center. When the button is pressed, current is briefly impeded by the back-EMF created by the coil, and is diverted through D1, making it flash briefly. Then the coil's reactance diminishes, allowing the current to flow through the coil and bypass the LED. When the pushbutton is released, the coil's magnetic field collapses, and the consequent forward-EMF circulates through D2, causing it to flash briefly. Note that the polarity of back-EMF and forward-EMF are opposite, which is why the LEDs in the circuit are oriented with opposite polarities.

The 220Ω resistor should be rated at 1/4 watt minimum, and the button should not be held down for long, as the electrical resistance of the coil is relatively low. The LEDs ideally should be rated for a minimal forward current of no more than 5mA.

Electrical and Magnetic Polarity

Various mnemonics and images have been created to assist in memorizing the polarity or direction of the magnetic field that will be created by a flow of electricity. The *right-hand rule* suggests that if the fingers of the right hand are curled around a coil in the same direction in which the turns of the coil were wound, and if conventional DC current also flows in this direction, the extended thumb will point in the direction of the principal force that can be created by the magnetic field.

By convention, the magnetic field is oriented from *south* to *north*, which can be remembered since the north end of the magnetic field will be

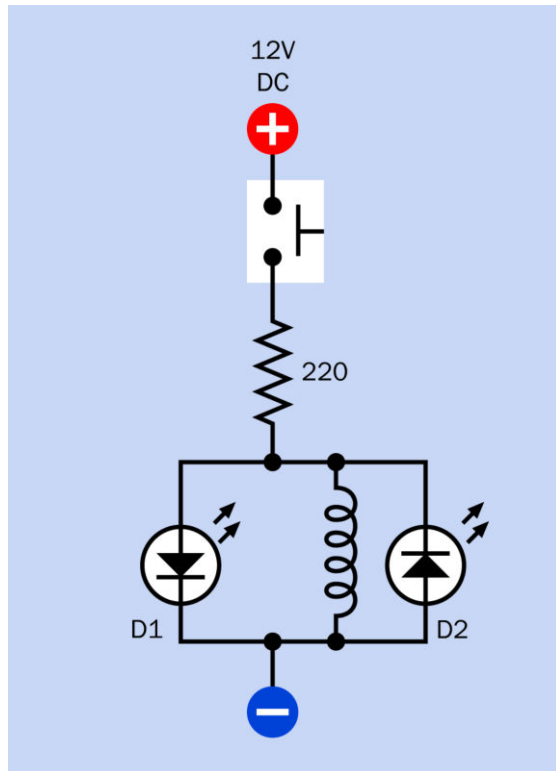


Figure 14-9. A test circuit to demonstrate the EMF and back-EMF created when DC current starts and stops passing through a coil. See text for details.

the negative end of the coil (north and negative both beginning with letter N). This mnemonic only works if conventional (positive) current flows through a coil that is wound clockwise.

Another model is the “corkscrew rule” in which we imagine conventional DC current flowing from the handle of a corkscrew, down through its metal section, toward the pointed end. If the corkscrew is turned clockwise, in the same direction as the electricity, the corkscrew will sink into the cork in the same as the direction as the resulting magnetic force.

Variants

Variants include core materials, core shapes, termination style (for through-hole mounting in

perforated board, or for surface-mount), and external finish (some inductors are dipped in insulating material, while others allow their copper magnet wire to be exposed).

In addition there are two functional variants: variable inductors and ferrite beads. Their schematic symbols are shown in [Figure 14-10](#).

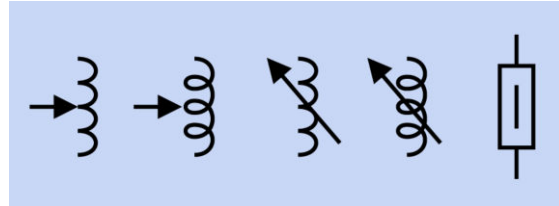


Figure 14-10. Schematic symbols for a ferrite bead (farthest right) and variable inductors (all other symbols, which are functionally identical).

Magnetic Cores

A magnetic core may be made from solid iron, plates of iron or steel separated by thin insulating material, powdered iron mixed with a binder, or a ferrite compound derived from nickel, zinc, manganese, or a combination. An iron core has at least 1,000 times the permeability of air, while some ferrites are 10,000 times as permeable.

One major disadvantage of a magnetic core is *hysteresis*, which in this context refers to the tendency of the core to retain some magnetic “memory” as a cycle of alternating current changes from positive to negative. This residual magnetism must be overcome by the next positive pulse of AC. The tendency of the core to retain magnetic polarity is known as its *retentivity*. Iron cores are especially retentive.

Another disadvantage of some magnetic cores is that they may host eddy currents induced by the magnetic field of the coil. These electrical currents tend to circulate through the core, reducing efficiency by generating waste heat, especially if coil currents are high. Forming a core from iron or steel plates, separated by thin layers of insulation, will inhibit these currents. Powdered iron

inhibits eddy currents because the particles have limited contact. Ferrites are nonconductive, and are therefore immune to eddy currents. They are widely used.

Hysteresis and eddy currents both incur energy losses with each AC cycle. Therefore, the losses increase linearly as the AC frequency increases. Consequently, inductor cores that suffer either of these problems are not well-suited to high frequencies.

Nonmagnetic Cores

The problems associated with magnetic cores may be avoided by winding the coil around a nonmagnetic core that may be hollow, ceramic, or plastic. A hollow core is referred to as an *air core*. The permeability of ceramic and plastic cores is close to that of air.

An inductor with a nonmagnetic core will be immune to eddy currents and retentivity, but will have to be significantly larger than a magnetic-cored coil with comparable inductance. In the case of a very primitive radio receiver, such as a crystal set, the air-cored coil that selects a radio frequency may be several inches in diameter. A basic circuit diagram for a crystal set (so-called because it uses a diode containing a germanium crystal) is shown in Figure 14-11. The antenna, at top, receives signals broadcast from radio stations. The coil can be tapped (as indicated by the black dots) as a simple way to select different inductance values, blocking all but a narrow range of frequencies. The T-shaped white component at right is a high-impedance earphone. The diode blocks the lower half of the alternating current in a radio signal, and since the signal is *amplitude-modulated*, the earphone responds to variations in intensity in the signal and reproduces the sound encoded in it.

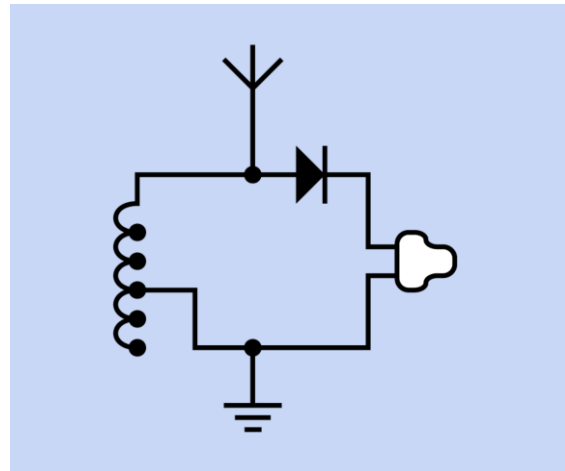


Figure 14-11. An early and basic application for an inductor is to select radio-station frequencies, as in this schematic for a crystal set. See text for details.

Variable Inductors

A *variable inductor*, also known as an *adjustable inductor*, is relatively uncommon but can be fabricated by using a magnetic core that penetrates the center of the inductor on an adjustable screw thread. The inductance of the assembly will increase as a larger proportion of the magnetic core penetrates into the open center of the coil. A photograph of a variable inductor is at Figure 14-12.

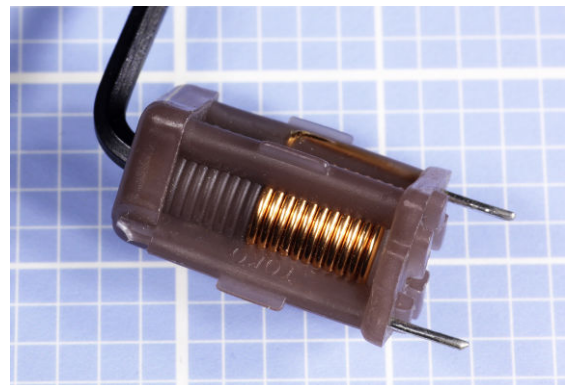


Figure 14-12. A variable inductor. Its inductance is adjusted via a screw thread that varies the insertion of the core in the coil. In this component the core is turned by inserting a hex wrench, as shown. It is rated from 0.09 μ H to 0.12 μ H.

Ferrite Beads

A *ferrite bead* inverts the design of a typical inductor by running a wire through a hole in the center of the bead, instead of coiling the wire around the core. Two ferrite beads are shown in [Figure 14-13](#). At top, the bead is divided into two sections, each mounted in one-half of a plastic clam shell, which can be closed around a wire. At bottom, the bead must be threaded onto a wire. The purpose is either to limit radio-frequency radiation from a wire by absorbing it into the bead (where it is transformed into heat), or to protect a wire from external sources of radio-frequency radiation. Computer cabling to external devices; lamp dimmers; and some types of motors can be sources of radio frequency.



Figure 14-13. Two examples of ferrite beads. They can inhibit radio-frequency radiation from a wire, or protect the wire from interference.

Toroidal Cores

The magnetic circuit created by a rod-shaped core must be completed by the lines of force traveling back around from one end of the rod to the other, through the surrounding air. Since air has low permeability, this is a major source of in-

efficiency. By comparison, a torus (a geometrical shape resembling a donut) completes the entire magnetic circuit inside its core. This significantly increases its efficiency. Also, because its field is better contained, a toroidal inductor needs little or no shielding to protect other components from stray magnetic effects.

Two through-hole toroidal inductors are shown in [Figure 14-2](#). Bottom left: Rated at $345\mu\text{H}$. Bottom right: Rated at $15\mu\text{H}$. The one at bottom-left has pins beneath it for insertion into a printed circuit board.

Surface-mount inductors often are toroidal to maximize the efficiency of a component that has to function on a very small scale. Examples are shown in [Figure 14-14](#), [Figure 14-15](#), and [Figure 14-16](#).

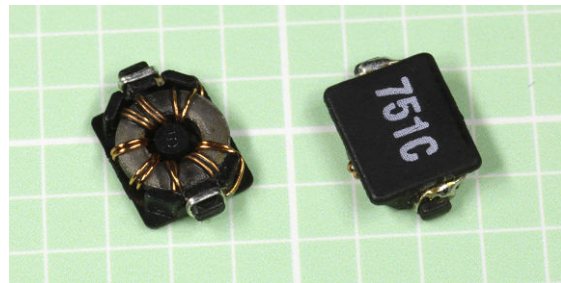


Figure 14-14. In a typical toroidal inductor, the coil is wrapped around a magnetic core shaped as a torus. This surface-mount component (viewed from the bottom, at left, and from the top, at right) is at the low end of the range of component sizes. It is rated at 750nH .

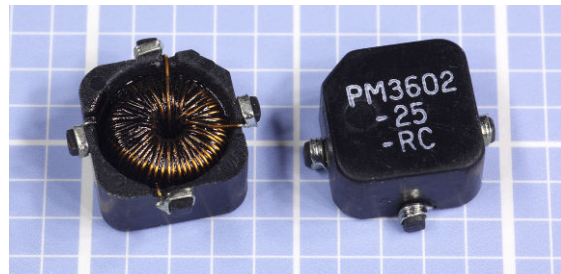


Figure 14-15. A medium-sized surface-mount toroidal inductor (viewed from the bottom, at left, and from the top, at right). It is rated at $25\mu\text{H}$.

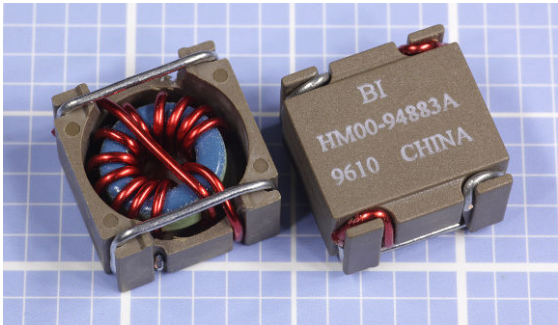


Figure 14-16. A larger-sized surface-mount toroidal inductor (viewed from the bottom, at left, and from the top, at right). It is rated at 3.8μH.

A chart showing some types of inductor cores, their commonly available inductances, and their maximum frequencies is shown in Figure 14-17.

Gyrator

A *gyrator* is a small network, sometimes encapsulated in a silicon chip, using **resistors**, a semiconductor, and a **capacitor** to simulate some but not all of the behavior of a coil-based inductor. The semiconductor may be a transistor or a capacitor, depending on the specific circuit. A sample schematic is shown in Figure 14-18. Because no magnetic effects are induced, the gyrator is completely free from the problems of saturation and hysteresis, which affect coils with cores, and also produces no back-EMF. It simply attenuates a signal initially, and then gradually lowers its reactance, thus imitating this aspect of an inductor.

A gyrator may be used where a coil may be unacceptable large (as in a cellular phone) or where signal quality is of paramount importance—for example, in a graphic equalizer or other audio components that perform signal processing at input stages, such as preamplifiers.

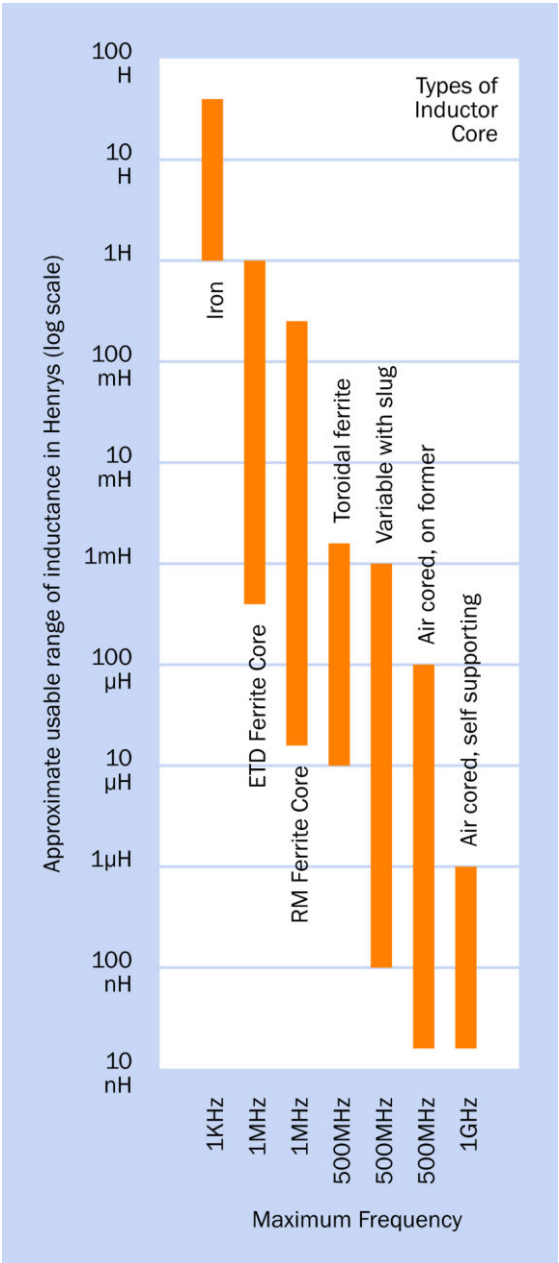


Figure 14-17. Some commonly used inductor cores and their characteristics. Adapted from "Producing wound components" by R.Clark@surrey.ac.uk.

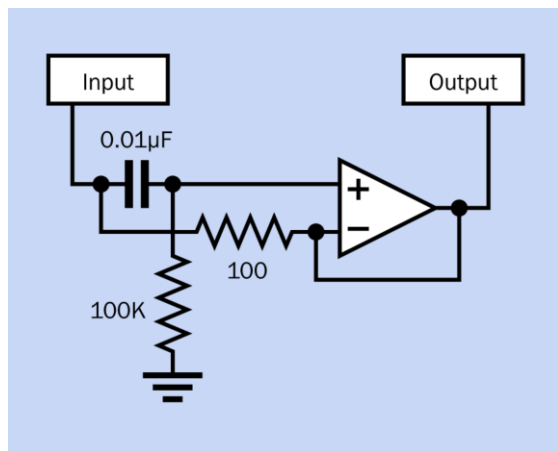


Figure 14-18. A possible schematic for a coil substitute known as a *gyrator*, which may be used where a conventional coil would be unacceptably bulky.

A gyrator does impose some limits on circuit design. While neither side of a real inductor needs to be at ground potential, a gyrator does require a ground connection. However, the performance advantages of gyrators are significant, as they can emulate high inductance without parasitic effects, can be more accurately calibrated (leading to more predictable performance), and do not create magnetic fields that can interfere with other components.

Values

Calculating Inductance

The magnetic inductance of a coil is measured with a unit known as the Henry, named after Joseph Henry, a pioneer in electromagnetism. It is defined by imagining a coil in which current is fluctuating, causing the creation of EMF. If the rate of fluctuation is 1 amp per second and the induced EMF is 1 volt, the inductance of the coil is 1 Henry.

The letter L is commonly used to represent inductance. To derive a useful formula, L will be expressed in microhenrys. If D is the diameter of a coil, N is the number of turns of wire, and W is

the width of coil (when the windings are viewed from the side, as shown in Figure 14-19), the precise relationship of the variables is complex but can be reduced to an approximate formula:

$$L = \text{approx } (D^2 * N^2) / 18 * D + (40 * W$$

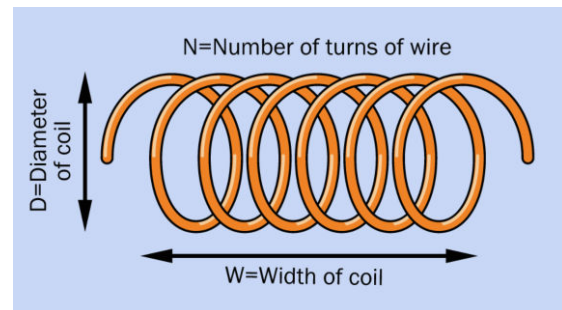


Figure 14-19. Dimensions of a coil, referenced by a formula to calculate its approximate inductance. See text for details.

From this, it is clear that inductance tends to increase with coil diameter, and also increases (more significantly) with the square of the number of turns. If the number of turns remains constant, inductance will be higher for a coil that is short and fat than for a coil that is narrow and long.

Because the Henry is a large unit, inductors in electronics circuits typically have their inductances measured in millihenrys (mH), microhenrys (μH), and nanohenrys (nH), where 1H = 1,000mH, 1 mH = 1,000μH, and 1μH = 1,000 nH. This relationship is shown in Figure 14-20.

Calculating Reactance

The reactance of an inductor (that is, its dynamic resistance to alternating current) varies with the frequency of the current. If f is the AC frequency (in Hertz), and L is the inductance (in Henrys), the reactance, X_L in ohms, is given by the formula:

$$X_L = 2 * \pi * f * L$$

From this equation, it's apparent that as the frequency tends toward zero (DC current), or if the inductance tends toward zero (a short piece of

nH	μH	mH
1	0.001	0.000001
10	0.01	0.00001
100	0.1	0.0001
1,000	1	0.001
10,000	10	0.01
100,000	100	0.1
1,000,000	1,000	1

Figure 14-20. Inductance is typically measured in nanohenrys (nH), microhenrys (μH), and millihenrys (mH). Equivalent values in these units are shown here.

straight wire), the reactance will tend toward zero. Conversely, the inductor will impede current increasingly as the frequency and/or the inductance increases.

Calculating Reluctance

The letter S is often used to represent reluctance, while Greek letter μ customarily represents permeability (not to be confused with the use of μ as a multiplication factor of 1/1,000,000, as in μF, meaning “microfarad”). If A is the area of cross-section of the magnetic circuit and L is its length:

$S = L / \mu * A$

Datasheet Terminology

A typical manufacturer’s datasheet should include an *inductance index* for an inductor, expressed in μH per 100 turns of wire (assuming the wire is in a single layer) for inductors with a powdered iron core, and mH per 1,000 turns of wire for inductors with ferrite cores.

The *DCR* is the DC resistance of an inductor, derived purely from the wire diameter and its length.

The *SRF* is the *self-resonant frequency*. An inductor should be chosen so that AC current passing through it will never get close to that frequency.

ISAT (or *I_{sat}*) is the saturation current, which results in a magnetic core losing its function as a result of magnetic saturation. When this occurs, inductance drops and the charge current rate increases drastically.

Series and Parallel Configurations

Because the inductance of a coil conducting DC current is proportional to the current, the calculations to derive the total inductance of coils in series or in parallel are identical to the calculations used for resistors.

In series, all the coils inevitably pass the same current, and the total inductance is therefore found by summing the individual inductances. When coils are wired in parallel, the current distributes itself according to the inductances; therefore, if L1 is the reluctance of the first coil, L2 is the reluctance of the second coil, and so on, the total reluctance L of the network is found from the formula:

$1/L = 1/L1 + 1/L2 + 1/L3. . .$

This is shown in *Figure 14-21*. In reality, differences between the coils (such as their electrical resistance), and magnetic interaction between the coils, will complicate this simple relationship.

Time Constant

Just as the *time constant* of a capacitor defines the rate at which it accumulates voltage when power is applied through a resistor, the time constant of an inductor defines the rate at which it gradually allows amperage to pass through it, overcoming the EMF generated by the coil. In both cases, the time constant is the number of seconds that the component requires to acquire approximately 63% of the difference between its current value and its maximum value. In the case of an inductor, suppose we assume zero internal resistance in the power source, zero resistance in

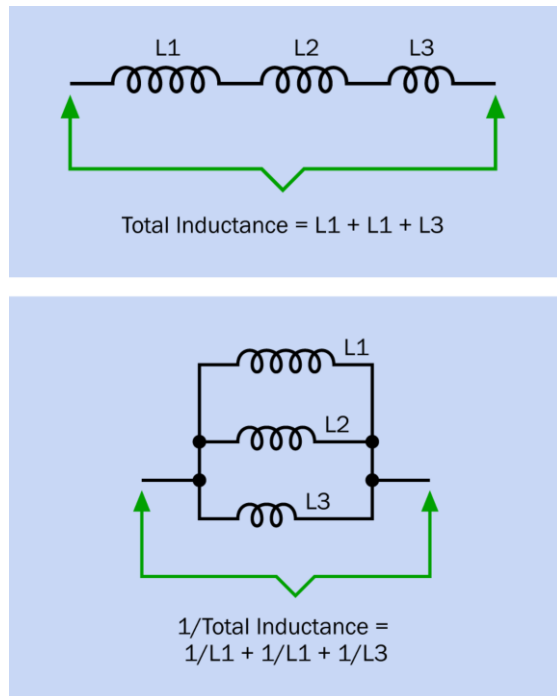


Figure 14-21. Calculating the total inductance of inductors in parallel (top) and series (bottom).

the coil windings, and an initial current of zero. If L is the inductance of the coil and R is the value of the series resistor, then the time constant— TC —is given in seconds by the formula

$$TC = L / R$$

Therefore a coil of 10 millihenrys (0.01 Henry) in series with a 100-ohm resistor will pass 63% of the full current in 0.0001 seconds, or 1/10 of a millisecond; it will take an equal additional amount of time for the current to rise by another 63% of the remaining difference between its charge and the maximum amperage of the circuit. In theory, the reactance of a coil can never diminish to zero, but in practice, five time constants are considered adequate to allow maximum current flow.

How to Use it

Because the inductance of an inductor peaks as current increases, and then gradually diminishes,

an inductor can be used to block or attenuate high frequencies. A circuit that does this is often referred to as a *low-pass filter*. The schematic and a graph suggesting its performance are shown in Figure 14-22. A basic application could be the *crossover network* in a loudspeaker system, where high-frequency signals are blocked from a low-frequency driver and are diverted to a high-frequency driver.

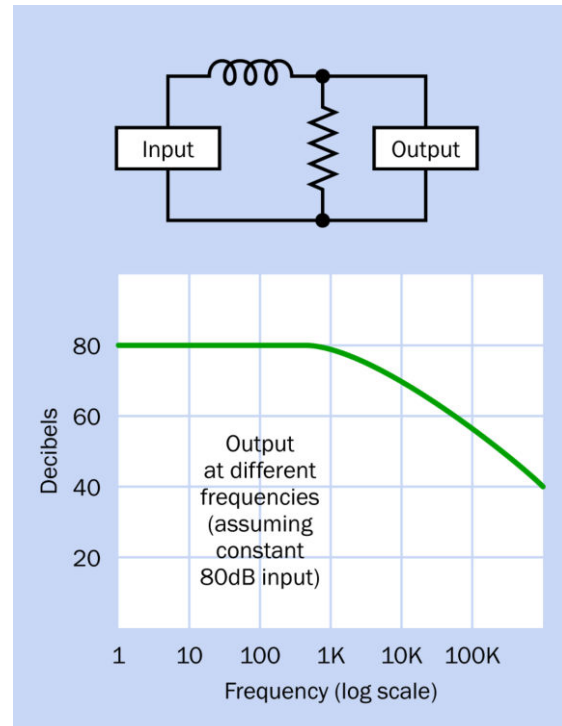


Figure 14-22. By using the ability of an inductor to block a range of frequencies, a low-pass filter blocks higher frequencies.

If the location of the inductor is shifted so that it shunts the signal away from the output, the results are reversed, and the circuit becomes a *high-pass filter*. The schematic and a graph suggesting its performance are shown in Figure 14-23.

Note that **capacitors** may also be used to create frequency filters, but because their function is roughly inverse to that of inductors, the place-

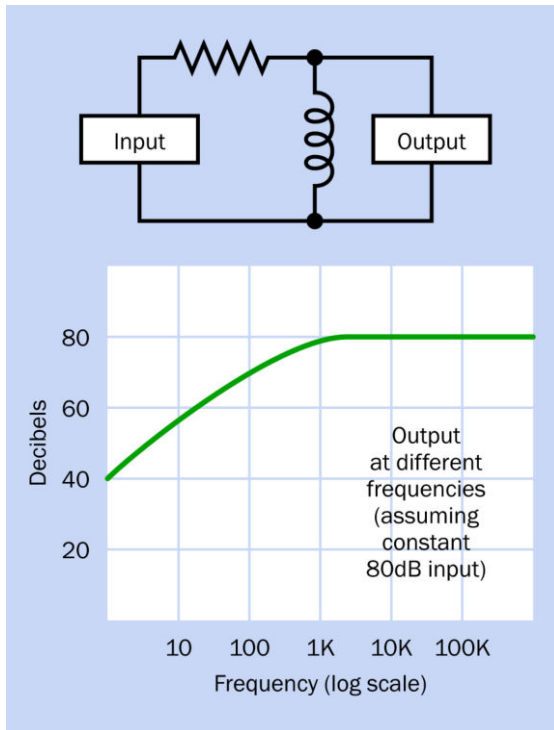


Figure 14-23. Here the inductor diverts low frequencies away from the output, allowing high frequencies to pass through.

ment of a capacitor in a circuit would be opposite to the placement of the inductor. Examples of filter circuits using capacitors are found in the entry for that component in this encyclopedia.

An inductor can be combined with a capacitor to form a *bandpass filter*, as shown in Figure 14-24. In this configuration, the inductor blocks the high frequencies while the capacitor blocks the low frequencies, allowing only a limited band of frequencies to get through.

Once again if the location of the components is shifted to shunt the signal away from the output, the results are reversed, as shown in Figure 14-25. This is known as a *notch filter*.

The performance of these filters will depend on the component values, and in most applications, additional components will be necessary. Sophisticated filter circuits are outside the scope of this encyclopedia.

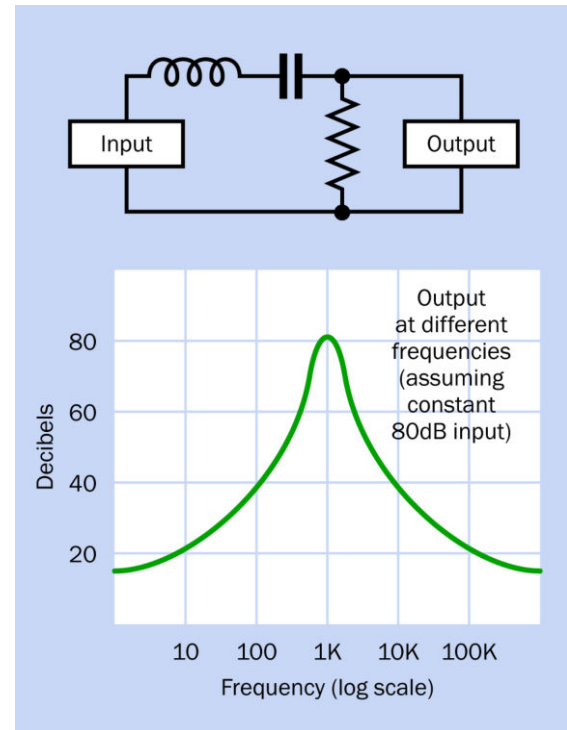


Figure 14-24. If the values of a capacitor and an inductor are correctly chosen, and the components are placed in series, the inductor blocks high frequencies while the capacitor blocks low frequencies, creating a bandpass filter, in which only a narrow band of frequencies can get through.

Inductors are of great importance in **DC-DC converters** and **AC-DC power supplies** where voltage changes are enabled by rapid switching. See the relevant entries of this encyclopedia for additional details.

Generally, as electronic equipment has become increasingly miniaturized, the unavoidable bulk of inductors has limited their application. However they may still be used to tune oscillators, to block sudden spikes in power supplies, and to protect equipment from sudden voltage spikes (they are used, for example, in surge suppressors for computing equipment).

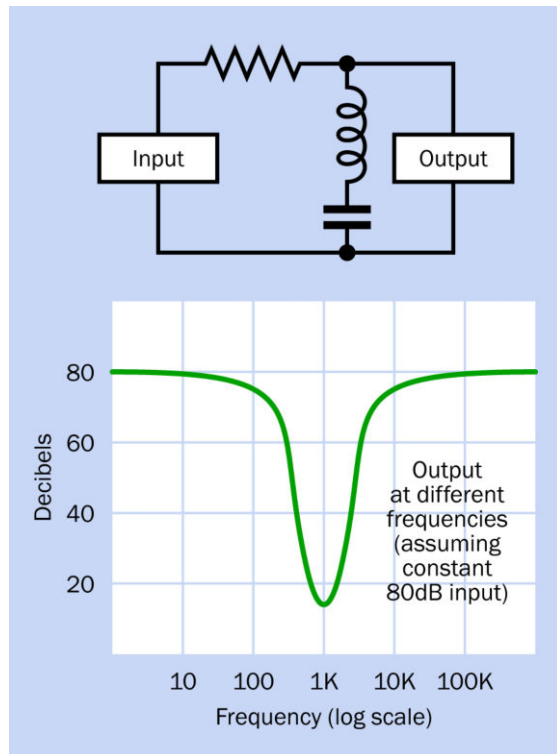


Figure 14-25. Here the capacitor and inductor block all frequencies except a narrow band, which they divert from the output. The result is a notch filter.

Core Choices

Air-cored inductors have relatively low inductance, because of their low permeability. However, they can be operated at very high frequencies up to the gigahertz range, and can tolerate higher peak currents.

Inductors with an iron core suffer increasing power losses due to hysteresis and eddy currents as the AC frequency passing through the inductor increases. Consequently, iron-cored inductors are not suitable for frequencies much above 10KHz.

Miniaturization

A low-value inductor can be formed by etching a spiral onto a circuit board, in applications where size must be minimized. They may also be incor-

porated in integrated circuit chips. However, in small devices such as cellular phones, it is more common to use a coil substitute such as a *gyrator*, as described previously.

What Can Go Wrong

Real-World Defects

The theoretically ideal inductor has no resistance or capacitance and suffers no energy losses. In reality, an inductor possesses both resistance and capacitance, also creates electrical noise, and may pick up electrical noise. It tends to create stray magnetic fields, and generally is more troublesome to deal with than its two cousins, the **resistor** and the **capacitor**.

Parasitic capacitance occurs between adjacent turns of wire. This capacitance becomes more significant at higher frequencies, leading ultimately to a situation where the coil becomes *self resonant*.

The workarounds for these problems involve coil geometries and choices of core material that go beyond the scope of this encyclopedia.

A *gyrator* should be considered as a possible substitute where inductors are troublesome or excessively expensive.

Saturation

Inductance increases as the current passing through a coil increases, but if a magnetic core is used, its contribution to inductance will stop abruptly when the core becomes magnetically *saturated*. In other words, when all of the randomly distributed *magnetic domains* in the core have been induced to align themselves with the pervasive magnetic field, the core cannot become more highly magnetized, and ceases to contribute to the inductance. Note that as a core approaches saturation levels, its hysteresis increases because reversing its magnetization requires greater energy. Antidotes to saturation would include a larger core, a lower current, a smaller number of turns in the coil, and using a core with lower permeability (such as air).

RF Problems

Radio frequencies (RF) introduce various problems affecting the efficiency of inductors. The *skin effect* is the tendency of high-frequency AC current to flow primarily on the surface of a strand of wire. The *proximity effect* refers to the tendency of the magnetic fields caused by adjacent wires to introduce eddy currents in the coil.

Both of these effects increase the effective resistance of the coil. Various coil geometries have been developed to minimize these effects, but are outside the scope of this encyclopedia. The fundamental lesson is that coils specifically designed for RF are the only ones that should be used with RF.

AC-AC transformer

OTHER RELATED COMPONENTS

- **AC-DC power supply** (See [Chapter 16](#))
- **DC-DC converter** (See [Chapter 17](#))
- **DC-AC inverter** (See [Chapter 18](#))

What It Does

A transformer requires an input of *alternating current* (AC). It transforms the input voltage to one or more output voltages that can be higher or lower.

Transformers range in size from tiny impedance-matching units in audio equipment such as microphones, to multi-ton behemoths that supply high voltage through the national power grids. Almost all electronic equipment that is designed to be powered by municipal AC in homes or businesses requires the inclusion of a transformer.

Two small power transformers are shown in [Figure 15-1](#). The one at the rear is rated to provide 36VAC at 0.8A when connected with a source of 125VAC. At front, the miniature transformer is a Radio Shack product designed to provide approximately 12VAC at 300mA, although its voltage will be more than 16VAC when it is not passing current through a load.

Transformer schematic symbols are shown in [Figure 15-2](#). The different coil styles at left and right are functionally identical. Top: A transformer with a magnetic core—a core that can be magnetized. Bottom: A transformer with an air core. (This type of transformer is rare, as it tends to be less efficient.) The input for the transformer is almost always assumed to be on the left, through the *primary* coil, while the output is on



Figure 15-1. Two small power transformers. The one at the rear measures approximately 1" × 2" × 2" and is rated to provide 36VAC at 0.8A. The term "sec" on the smaller unit is an abbreviation for "secondary," referring to the rating for its secondary winding.

the right, through the *secondary* coil. Often the two coils will show differing numbers of turns to indicate whether the transformer is delivering a reduced voltage (in which case there will be fewer turns in the secondary coil) or an increased voltage (in which case there will be fewer turns in the primary coil).

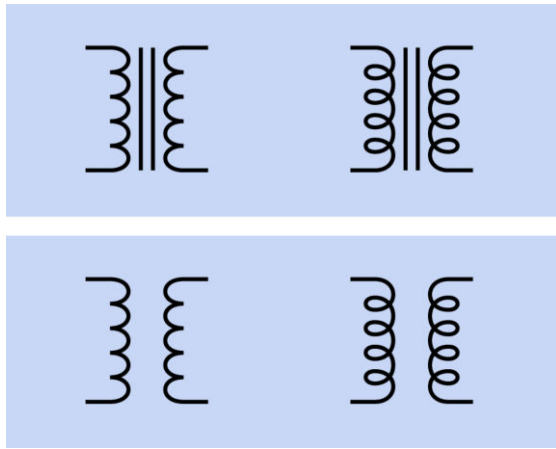


Figure 15-2. Alternate symbols for a transformer with a ferromagnetic core (top) and air core (bottom). The differing coil symbols at left and right are functionally identical.

How It Works

A simplified view of a transformer is shown in [Figure 15-3](#). Alternating current flowing through the primary winding (orange) induces magnetic flux in a laminated core formed from multiple steel plates. The changing flux induces current in the secondary winding (green), which provides the output from the transformer. (In reality, the windings usually consist of thousands of turns of thin magnet wire, also known as enameled wire; and various different core configurations are used.)

The process is known as [mutual induction](#). If a load is applied across the secondary winding, it will draw current from the primary winding, even though there is no electrical connection between them.

In an ideal, lossless transformer, the ratio of turns between the two windings determines whether the output voltage is higher, lower, or the same as the input voltage. If V_p and V_s are the voltages across the primary and secondary windings respectively, and N_p and N_s are the number of turns of wire in the primary and secondary windings, their relationship is given by this formula:

$$V_p / V_s = N_p / N_s$$

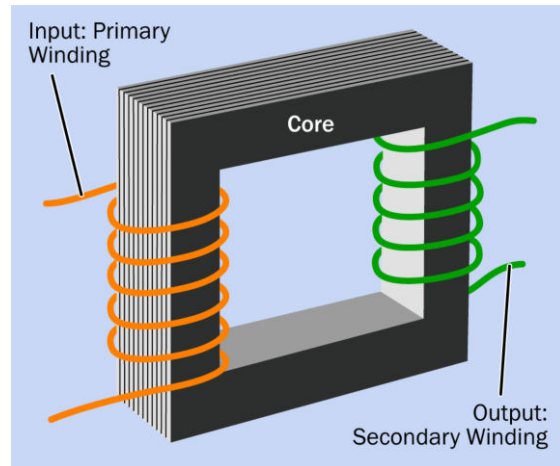


Figure 15-3. Three basic parts of a transformer, shown in simplified form.

A simple rule to remember is that fewer turns = lower voltage while more turns = higher voltage.

A [step-up transformer](#) has a higher voltage at its output than at its input, while a [step-down transformer](#) has a higher voltage at its input than at its output. See [Figure 15-4](#).

In an ideal, lossless transformer, the power input would be equal to the power output. If V_{in} and V_{out} are the input and output voltages, and I_{in} and I_{out} are the input and output currents, their relationship is given by this formula:

$$V_{in} * I_{in} = V_{out} * I_{out}$$

Therefore, if the transformer doubles the voltage, it allows only half as much current to be drawn from the secondary winding; and if the voltage is cut in half, the available current will double.

Transformers are not 100% efficient, but they can be more than 98% efficient, and relationships between voltage, current, and the number of turns in the windings are reasonably realistic.

When the transformer is not loaded, the primary winding behaves like a simple inductor with reactance that inhibits the flow of current. Therefore a power transformer will consume relatively

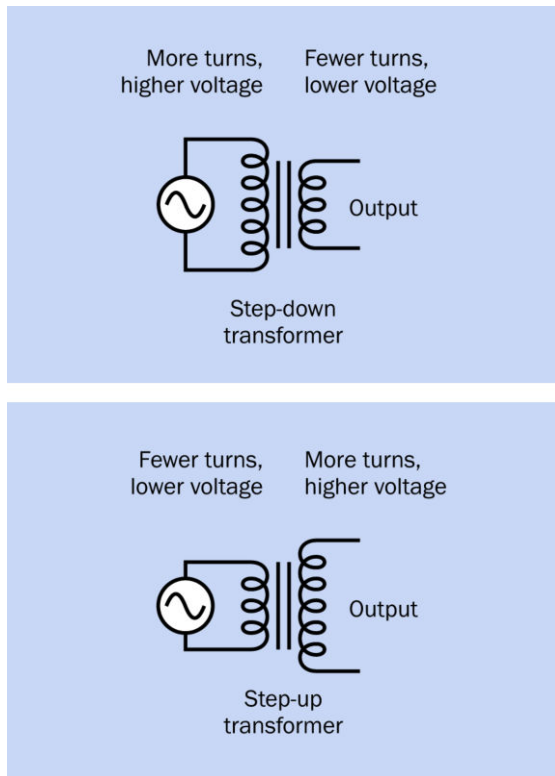


Figure 15-4. The ratio of input voltage to output voltage is equal to the ratio of primary turns to secondary turns in the transformer windings, assuming a transformer of 100% efficiency.

little electricity if it is left plugged in to an electrical outlet without any load connected to its output side. The power that it does consume will be wasted as heat.

The Core

The ferromagnetic core is often described as being made of iron, but in reality is more often fabricated from high permeability silicon steel. To reduce losses caused by eddy currents, the core is usually laminated—assembled from a stack of plates separated from each other by thin layers of varnish or a similar insulator. Eddy currents tend to be constrained within the thickness of each plate.

Because a DC voltage would cause magnetic saturation of the core, all transformers must operate

with alternating current or pulses of current. The windings and geometry of a transformer are optimized for the frequency range, voltage, and current at which it is designed to operate. Deviating significantly from these values can damage the transformer.

Taps

A *tap* on a transformer is a connection part-way through the primary or (more often) the secondary coil. On the primary side, applying an input between the start of a coil and a tap part-way through the coil will reduce the number of turns to which the voltage is applied, therefore increasing the ratio of output turns to input turns, and increasing the output voltage. On the secondary side, taking an output between the start of a coil and a tap part-way through the coil will reduce the number of turns from which the voltage is taken, therefore decreasing the ratio of output turns to input turns, and decreasing the output voltage. This can be summarized:

- A tap on the primary side can increase output voltage.
- A tap on the secondary side can provide a decreased output voltage.

In international power adapters, a choice of input voltages may be allowed by using a double-throw switch to select either the whole primary winding, or a tapped subsection of the winding. See [Figure 15-5](#). Modern electronics equipment often does not require a voltage adapter, because a **voltage regulator** or **DC-DC converter** inside the equipment will tolerate a wide range of input voltages while providing a relatively constant output voltage.

A transformer's secondary winding is often tapped to provide a choice of output voltages. In fact, most power transformers have at least two outputs, since the cost of adding taps to the secondary winding is relatively small. As an

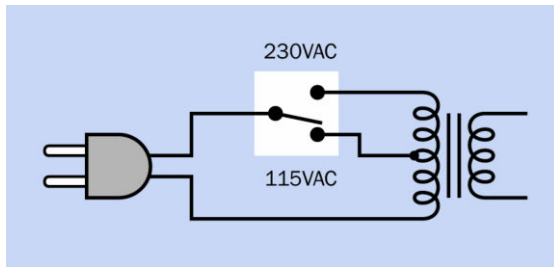


Figure 15-5. An international power adapter can provide a fixed output voltage by using a double-throw switch to apply 230VAC voltage across a transformer's primary winding, or 115VAC to a tapped midpoint of the primary winding.

alternative to tapped outputs, two or more separate secondary windings may be used, allowing the outputs to be electrically isolated from each other. See [Figure 15-6](#).

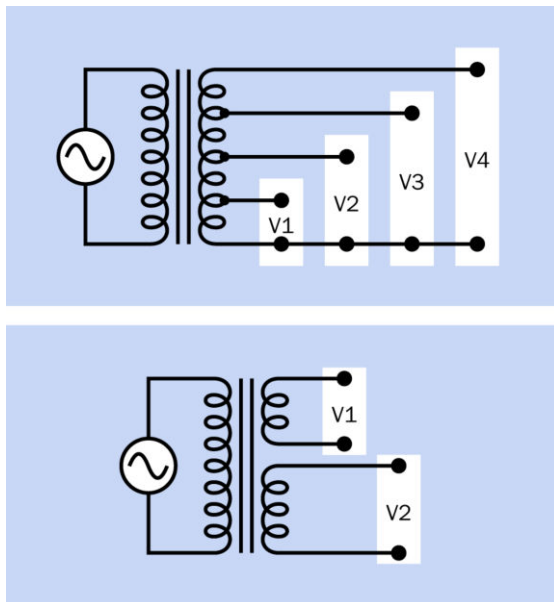


Figure 15-6. Multiple output voltages may be obtained from a transformer by tapping into the secondary winding (top) or using two or more separate secondary windings (bottom), in which case the outputs will be electrically isolated from each other.

If the winding on the primary side of a transformer is coiled in the same direction as the winding on the secondary side, the output voltage will be 180 degrees out of phase with the

input voltage. In schematics, a dot is often placed at one end of a transformer coil to indicate where the coil begins. If the dots on the primary and secondary sides are at the same ends of the coils, there will be a 180 degree phase difference between input and output. For many applications (especially where the output from a power transformer is going to be converted to DC), this is immaterial.

If there is a center tap on the secondary winding, and it will be referenced as ground, the voltages relative to it, at opposite ends of the secondary winding, will be out of phase. See [Figure 15-7](#).

Variants

Core Shapes

The *shell core* is a closed rectangle, as shown in [Figure 15-3](#). This is the most efficient but most costly to manufacture. A C-shaped core is another option (three sides of the rectangle) and an E-I core is popular, consisting of a stack of E-shaped plates with two coils wound around the top and bottom legs of the E, or wound concentrically around the center leg of the E. An additional stack of straight plates is added to close the gaps in the E and form a magnetic circuit.

In [Figure 15-8](#), the small transformer from [Figure 15-1](#) has been sliced open with a band saw and a belt sander to reveal a cross-section of its windings. This clearly shows that its primary and secondary windings are concentric. It also reveals the configuration of its core, which is in the E-I format. In [Figure 15-9](#), the E-I configuration is highlighted to show it more clearly.

Power Transformer

Typically designed to be bolted onto a chassis or secured inside the case or cabinet housing a piece of electrical equipment with solder tabs or connectors allowing wires to connect the transformer to the power cord, on one side, and a cir-

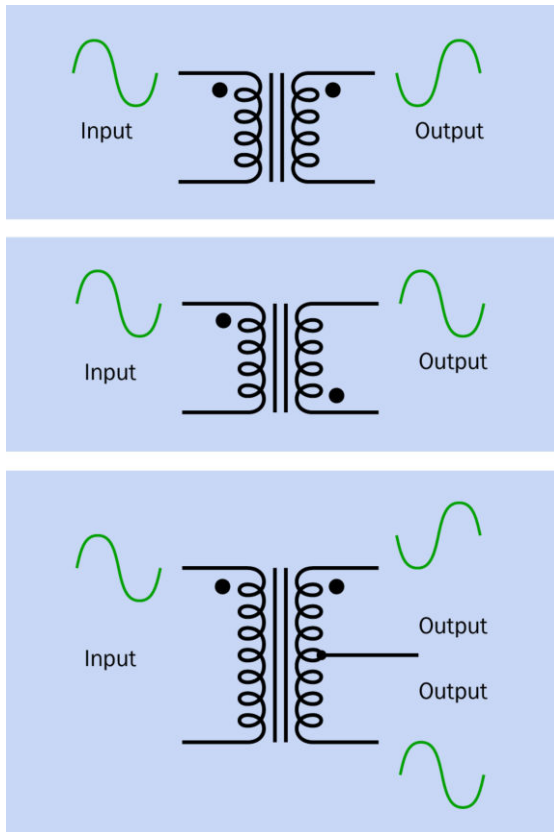


Figure 15-7. A dot indicates the start of each winding. Where primary and secondary windings are in the same direction, the voltage output will be 180 degrees out of phase with the input. Where the dots indicate windings in opposite directions, the voltage output will have the same phase as the input. Where a center tap on the secondary winding serves as a common ground, the voltages at opposite ends of the secondary winding will be opposite to each other in phase.

cuit board, on the other side. Smaller power transformers such as the one in [Figure 15-1](#) have “through-hole” design with pins allowing them to be inserted directly onto circuit boards.

Plug-in Transformer

Usually sealed in a plastic housing that can be plugged directly into a wall power outlet. They are visually identical to **AC adapters** but have an AC output instead of a DC output.

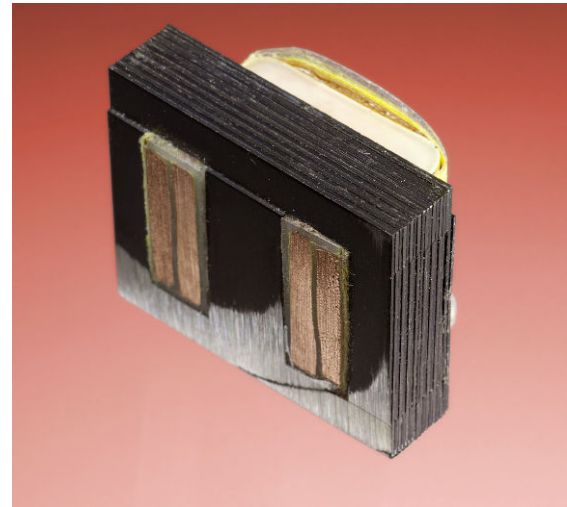


Figure 15-8. The small transformer from the first figure in this entry is shown sliced open to reveal its internal configuration.

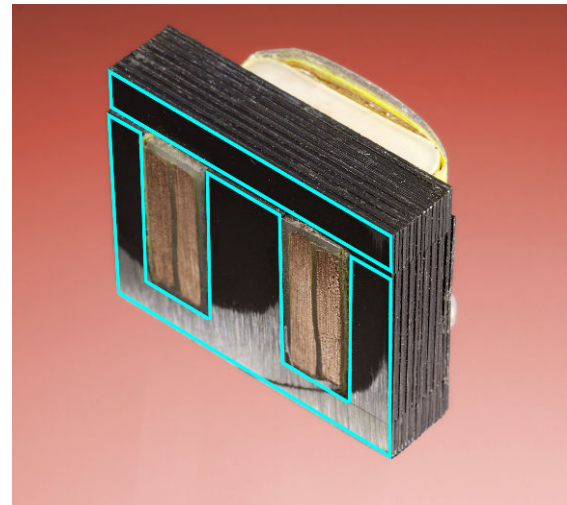


Figure 15-9. The “EI” shaped plates that form the core of the transformer are outlined to show their edges.

Isolation Transformer

Also known as a **1:1 transformer** because it has a 1:1 ratio between primary and secondary windings, so that the output voltage will be the same as the input voltage. When electrical equipment is plugged into the isolation transformer, it is separated from the electrical ground of AC power wiring. This reduces risk when working on

“live” equipment, as there will be negligible electrical potential between itself and ground. Consequently, touching a grounded object while also touching a live wire in the equipment should not result in potentially lethal current passing through the body.

Autotransformer

This variant uses only one coil that is tapped to provide output voltage. Mutual induction occurs between the sections of the coil. An autotransformer entails a common connection between its input and output, unlike a two-coil transformer, which allows the output to be electrically isolated from the input. See [Figure 15-10](#). Autotransformers are often used for impedance matching in audio circuits, and to provide output voltages that differ only slightly from input voltages.

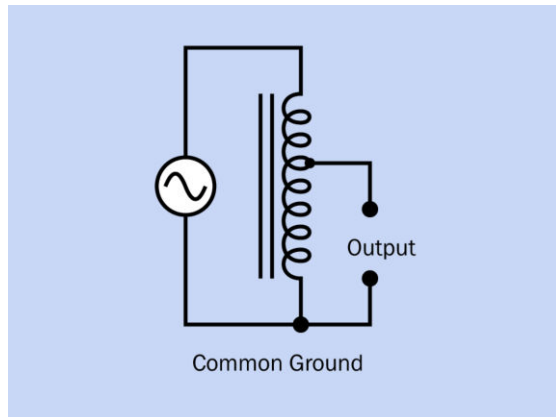


Figure 15-10. An autotransformer contains only one coil and core. A reduced output voltage can be obtained by tapping into the coil. A common connection prevents the output from being electrically isolated from the input.

Variable Transformer

A variable transformer, also known as a *variac*, resembles a wire-wound **potentiometer**. Only one winding is used. A *wiper* can be turned to contact the winding at any point, and serves as a movable tap. Like an autotransformer, a variable transformer entails a common connection between input and output.

Audio Transformer

When a signal is transmitted between two stages of a circuit that have different impedance, the signal may be partially reflected or attenuated. (Impedance is measured in ohms but is different from DC electrical resistance because it takes into account reactance and capacitance. It therefore varies with frequency.)

A device of low input impedance will try to draw significant current from a source, and if the source has high output impedance, its voltage will drop significantly as a result. Generally, the input impedance of a device should be at least 10 times the output impedance of the device that is trying to drive it. Passive components (resistors, and/or capacitors, and/or coils) can be used for impedance matching, but in some situations a small transformer is preferable.

If N_p and N_s are the number of turns of wire in the transformer primary and secondary windings, and Z_p is the impedance of a device (such as an audio amplifier) driving the transformer on its primary side, and Z_s is the impedance of a device (such as a loudspeaker) receiving power from the secondary side:

$$N_p / N_s = \sqrt{Z_p / Z_s}$$

Suppose that an audio amplifier with rated output impedance of 640Ω is driving a loudspeaker with 8Ω impedance. A matching transformer would be chosen with a ratio of primary turns to secondary turns give by:

$$\sqrt{(640/8)} = \sqrt{80} = \text{approximately } 9:1$$

The two transformers in [Figure 15-11](#) are through-hole components designed for telecommunications purposes, but are capable of passing audio frequencies and can be used for impedance matching in applications such as a preamplifier.

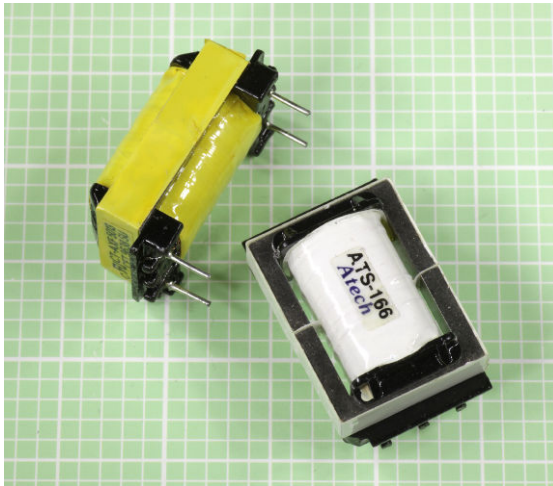


Figure 15-11. Through-hole transformers. See text for details.

In [Figure 15-12](#), the transformers are designed for audio coupling. The one on the right has impedances of 500 ohms (primary) and 8 ohms (secondary). On the left is a fully encapsulated line matching transformer with a 1:1 turns ratio.

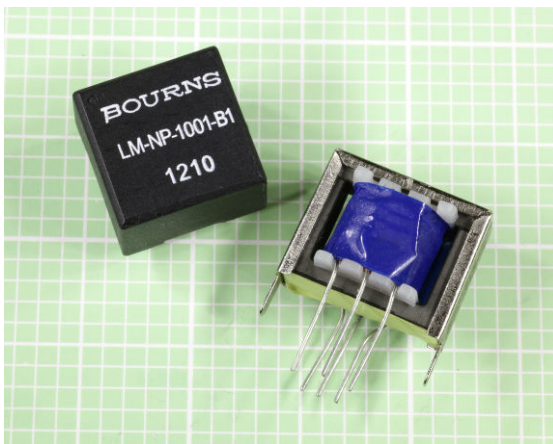


Figure 15-12. Through-hole transformers. See text for details.

Split-Bobbin Transformer

This variant has primary and secondary coils mounted side by side to minimize capacitive coupling.

Surface-Mount Transformer

May be less than 0.2" square and is used for impedance matching, line coupling, and filtering. Two surface-mount transformers are shown in [Figure 15-13](#).

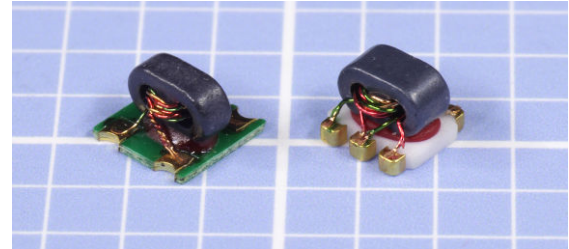


Figure 15-13. Two surface-mount transformers, each measuring less than 0.2" square, typically used in communications equipment and suited for frequencies higher than 5 MHz.

Values

When selecting a power transformer, its power handling capability is the value of primary interest. It is properly expressed by the term VA, derived from "volts times amps." VA should not be confused with watts because watts are measured instantaneously in a DC circuit, whereas in an AC circuit, voltage and current are fluctuating constantly. VA is actually the *apparent power*, taking reactance into account.

The relationship between VA and watts will vary depending on the device under consideration. In a worst-case scenario:

$$W = 0.65 \text{ VA (approximately)}$$

In other words, the averaged power you can draw from a transformer should be no less than two-thirds of its VA value.

Transformer specifications often include input voltage, output voltage, and weight of the component, all of which are self-explanatory. Coupling transformers may also specify input and output impedances.

How to Use it

For most electronic circuits, a power transformer will be followed by a [rectifier](#) to convert AC to DC, and **capacitors** to smooth fluctuations in the supply. Using a prepackaged **power supply** or [AC adapter](#) that already contains all the necessary components will be more time-effective and probably more cost-effective than building a power supply from the ground up. See [Chapter 16](#).

What Can Go Wrong

Reversal of Input and Output

Suppose a transformer is designed to provide an output voltage of 10 volts from domestic AC power of 115 volts. If the wrong side of the transformer is connected with 115VAC by mistake, the output will now be more than 1,000 volts—easily enough to cause death, quite apart from destroying components that are connected with it. Reversing the transformer in this way may also destroy it. Extreme caution is advisable when making connections with power transformers. A meter should be used to check output voltage. All devices containing transformers should be fused on the live side and grounded.

Shock Hazard from Common Ground

When working on equipment that uses an autotransformer, the chassis will be connected through the transformer to one side of 115VAC power. So long as a plug is used that prevents reversed polarity, the chassis should be “neutral.” However, if an inappropriate power cord is used, or if the power outlet has been wired incorrectly, the chassis can become live. For protection, before working on any device that uses 115VAC power with an autotransformer, plug the device into an isolation transformer, and plug the isolation transformer into the wall outlet.

Accidental DC Input

If DC current is applied to the input side of a transformer, the relatively low resistance of the primary coil will allow high current that can destroy the component. Transformers should only be used with alternating current.

Overload

If a transformer is overloaded, heat will be generated that may be sufficient to destroy the thin layers of insulation between coil windings. Consequently, input voltage can appear unexpectedly on the output side. Transformers with a toroidal (circular) core are especially hazardous in this respect, as their primary and secondary windings usually overlap.

Some (not all) power transformers contain a thermal fuse that melts when it exceeds a temperature threshold. If the fuse is destroyed, the transformer must be discarded.

The consequences of moderate overloading may not be obvious, and can be cumulative over time. Ventilation or heat sinkage should be taken into account when designing equipment around a power transformer.

Incorrect AC Frequency

Single-phase AC power in the United States fluctuates at 60Hz, but Great Britain and some other countries use AC power at 50Hz. Many power transformers are rated to be compatible with either frequency, but if a transformer is specifically designed for 60Hz, it may eventually fail by overheating if it is used with a 50Hz supply. (A 50Hz transformer can be used safely with 60Hz AC.)

AC-DC power supply

Also known as an *AC adapter*. When packaged as a palm-sized plastic package that plugs directly into a power outlet, it is occasionally known colloquially as a *wall-wart*.

OTHER RELATED COMPONENTS

- **transformer** (See [Chapter 15](#))
- **DC-DC converter** (See [Chapter 17](#))
- **DC-AC inverter** (See [Chapter 18](#))

What It Does

An AC-DC power supply converts *alternating current* (AC) into the *direct current* (DC) that most electronic devices require, usually at a lower voltage. Thus, despite its name, a power supply actually requires an external supply of power to operate.

Larger products, such as computers or stereo equipment, generally have a power supply contained within the device, enabling it to plug directly into a wall outlet. Smaller battery-powered devices, such as cellular phones or media players, generally use an external power supply in the form of a small plastic pod or box that plugs into a wall outlet and delivers DC via a wire terminating in a miniature connector. The external type of power supply is often, but not always, referred to as an *AC adapter*.

Although an AC-DC power supply is not a single component, it is often sold as a preassembled modular unit from component suppliers.

Variants

The two primary variants are a *linear regulated power supply* and *switching power supply*.

Linear Regulated Power Supply

A linear regulated power supply converts AC to DC in three stages:

1. A power **transformer** reduces the AC input to lower-voltage AC.
2. A *rectifier* converts the AC to unsmoothed DC. Rectifiers are discussed in the entry on **diodes** in this encyclopedia.
3. A **voltage regulator**, in conjunction with one or more **capacitors**, controls the DC voltage, smooths it, and removes *transients*. The regulator is properly known as a *linear voltage regulator* because it contains one or more **transistors**, which are functioning in linear mode—that is, responding linearly to fluctuations in base current, at less than their saturation level. The linear voltage regulator gives the linear regulated power supply its name.

A simplified schematic of a linear regulated power supply is shown in [Figure 16-1](#).

This type of power supply may be described as *transformer-based*, since its first stage consists of a transformer to drop the AC input voltage before it is rectified.

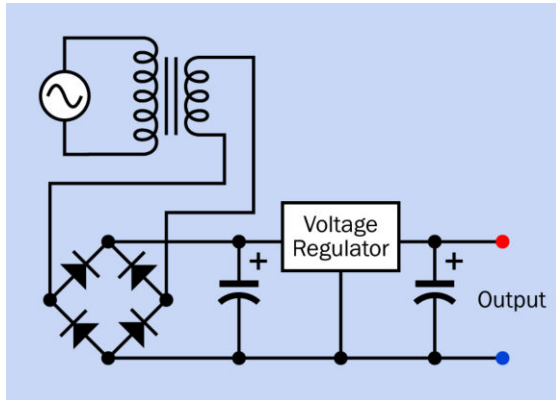


Figure 16-1. A basic linear regulated power supply.

Because the rectifier in a power supply generally passes each pulse of AC through a pair of silicon diodes, it will impose a voltage drop of about 1.2V at peak current. A smoothing capacitor will drop the voltage by about 3V as it removes ripple from the current, whereas a voltage regulator typically requires a difference of at least 2V between its input and its output. Bearing in mind also that the AC input voltage may fluctuate below its rated level, the output from the power transformer should be at least 8VAC higher than the ultimate desired DC output. This excess power will be dissipated as heat.

The basic principle of the linear regulated power supply originated in the early days of electronic devices such as radio receivers. A transistorized version of this type of power supply remained in widespread use through the 1990s. Switching power supplies then became an increasingly attractive option as the cost of semiconductors and their assembly decreased, and high-voltage transistors became available, allowing the circuit to run directly from rectified line voltage with no step-down power transformer required.

Some external AC adapters are still transformer-based, but are becoming a minority, easily identified by their relatively greater bulk and weight. An example is shown in [Figure 16-2](#).

[Figure 16-3](#) shows the handful of components inside a cheap, relatively old AC adapter. The out-



Figure 16-2. A simple transformer-based power supply can be encapsulated in a plastic shell, ready to plug into a power outlet. However, today this format more typically contains a switching power supply, which is usually lighter, smaller, and cheaper.

put from a power transformer is connected directly to four diodes (the small black cylinders), which are wired as a full-wave rectifier. A single electrolytic capacitor provides some smoothing, but because there is no voltage regulator, the output will vary widely depending on the load. This type of AC adapter is not suitable for powering any sensitive electronic equipment.

Switching Power Supply

Also known as a *switched-mode power supply*, an *SMPS*, or *switcher*, it converts AC to DC in two stages.

1. A *rectifier* changes the AC input to unsmoothed DC, without a power transformer.
2. A DC-DC **converter** switches the DC on and off at a very high frequency using *pulse-width modulation* to reduce its average effective voltage. Often the converter will be the *flyback* type, containing a transformer, but the high-frequency switching allows the transformer to be much smaller than the power transformer required in a linear regulated power supply. See the DC-DC **converter** entry in this encyclopedia for an explanation of the working principles.

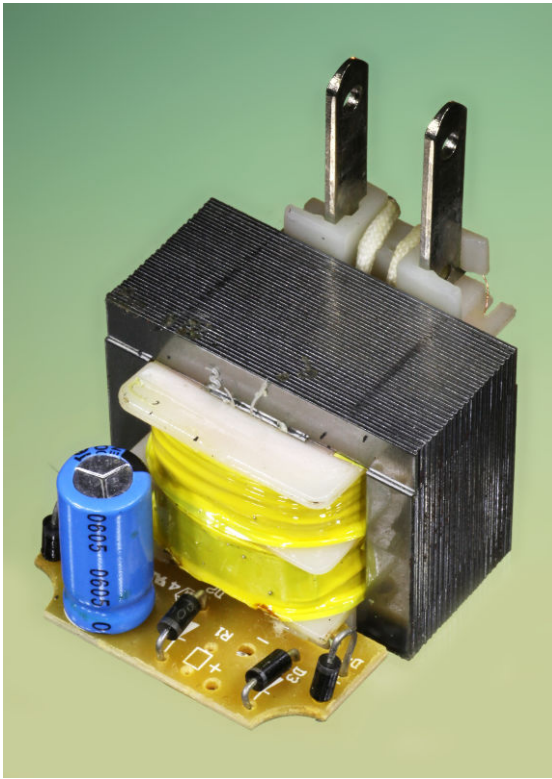


Figure 16-3. A relatively old, cheap AC adapter contains only the most rudimentary set of components, and does not supply the kind of properly regulated DC power required by electronic equipment.

A simplified schematic of a switching power supply is shown in Figure 16-4.

The interior of a relatively early switching power supply designed to deliver 12VDC at up to 4A is shown in Figure 16-5. This supply generated considerable waste heat, necessitating well-spaced components and a ventilated enclosure.

The type of small switching power supply that is now almost universally used to power laptop computers is shown in Figure 16-6. Note the smaller enclosure and the higher component count than in the older power supply shown in Figure 16-5. The modern unit also delivers considerably more power, and generates less waste heat. Although this example is rated at 5A, the

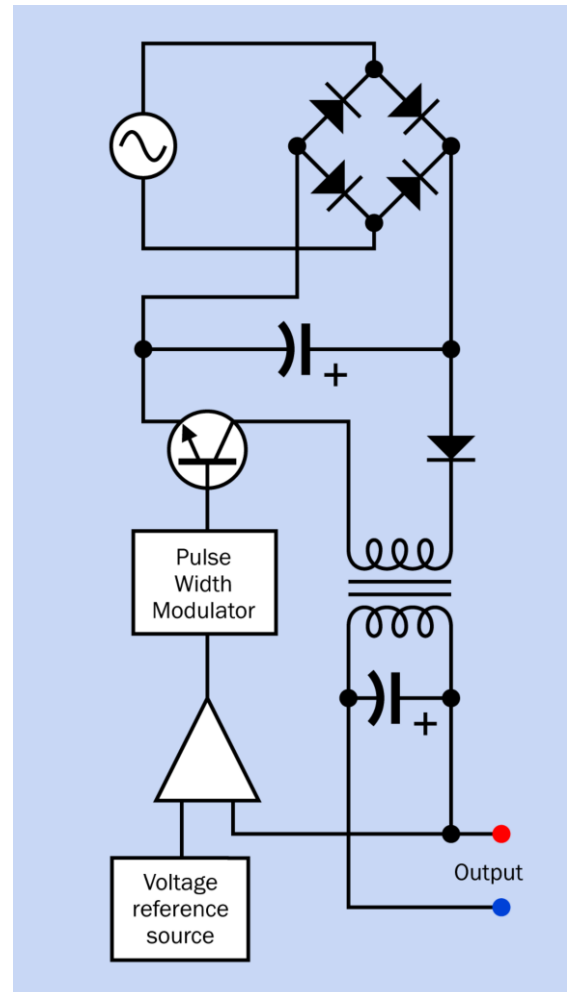


Figure 16-4. Greatly simplified schematic showing the principal components of a switching power supply. Note the absence of a 115VAC power transformer. The transformer that is inserted subsequently in the circuit functions in conjunction with the high switching frequency, which allows it to be very much smaller, cheaper, and lighter.

transformer (hidden under the yellow wrapper at the center of the unit) is smaller than the power transformer that would have been found in an old-style AC adapter delivering just 500mA.

The modern power supply is completely sealed, where earlier versions required ventilation. On



Figure 16-5. The interior of an early switching power supply.

the downside, the plastic case of the switching supply requires a metal liner (removed for this photograph) to contain high-frequency electromagnetic radiation.

Unregulated Power Supply

Typically this consists of a transformer and rectifying diodes with little or no smoothing or voltage control of the output.

Adjustable Power Supply

This is usually a linear power supply incorporating an adjustable **voltage regulator**. This type of supply has laboratory applications and is found as a benchtop item to power electronics design projects during their development.



Figure 16-6. The interior of the type of switching power supply that powers a laptop computer.

Voltage Multiplier

Devices such as photocopiers and laser printers, televisions, cathode-ray tubes, and microwave ovens require voltages significantly higher than those supplied by domestic AC power outlets. A voltage multiplier usually contains a step-up transformer followed by DC conversion components, but detailed consideration is outside the scope of this encyclopedia.

Formats

An *open frame* power supply consists of components on a circuit board, usually mounted on a metal chassis, with no enclosure or fan cooling.

A *covered* power supply is enclosed in a protective perforated metal box with a cooling fan if needed. Power supplies sold for desktop computers are usually in this format.

Power supplies are also available in rack-mount and DIN-rail formats.

How to Use it

Because a switching power supply contains no power transformer, it is lighter and smaller, and may be cheaper than a linear power supply. It is also more efficient and generates less waste heat. These advantages have made switching power supplies the most popular option to provide DC power for electronics devices. However, the high-frequency switching tends to create *electromagnetic interference (EMI)*, which must be filtered to protect the output of the device and also to minimize the risk of this interference feeding back into AC power wiring. The high-frequency switched power may also generate harmonics, which must be suppressed.

High-quality linear regulated power supplies still find application in laboratory equipment, low-noise signal processing, and other niches where excellent regulation and low-ripple output are necessary. They are relatively heavy, bulky, and inefficient.

See [Figure 16-7](#) for a chart comparing the advantages and disadvantages of linear and switching power supplies.

What Can Go Wrong

High Voltage Shock

One or more capacitors in a power supply may retain a relatively high voltage for some time after the unit has been unplugged. If the power supply is opened for inspection or repairs, caution is necessary when touching components.

Capacitor Failure

If *electrolytic capacitors* fail in a switching power supply (as a result of manufacturing defects, disuse, or age), allowing straight-through conduction of alternating current, the high-frequency switching semiconductor can also fail, allowing

input voltage to be coupled unexpectedly to the output. Capacitor failure is also a potential problem in linear power supplies. For additional information on capacitor failure modes, see [Chapter 12](#).

	Switching power supply	Linear power supply
Component count	High	Medium
Load regulation	0.05% to 0.5%	0.005% to 0.2%
Line regulation	0.05% to 0.2%	0.005% to 0.05%
Ripple (RMS)	10mV to 25mV	0.25mV to 1.5mV
Efficiency	70% to 85%	40% to 60%
EMI	High	Very low
Leakage	High	Low
Physical size	Small	Large
Weight	Light	Heavy
Heat management	Usually fan-cooled	Usually cooled by convection

Figure 16-7. Comparison of attributes of linear regulated power supplies and switching power supplies. (Adapted from Acopian Technical Company.)

Electrical Noise

If *electrolytic capacitors* are used, their gradual deterioration over time will permit more electrical noise associated with high-frequency switching in a switching power supply.

Peak Inrush

A switching power supply allows an initial inrush or surge of current as its capacitors accumulate their charge. This can affect other components in the circuit, and requires fusing that tolerates brief but large deviations from normal power consumption.

DC-DC converter

Often referred to as a *switching regulator*, and sometimes as a *switcher*, not to be confused with a *switching power supply*.

OTHER RELATED COMPONENTS

- **AC-DC power supply** (See [Chapter 16](#))
- **voltage regulator** (See [Chapter 19](#))
- **DC-AC inverter** (See [Chapter 18](#))

What It Does

A DC-DC converter, often referred to simply as a *converter*, receives a DC voltage as its input and converts it to a regulated DC voltage as its output. The output voltage may be higher or lower than the input voltage, may be user-adjustable by adding an external resistor, and may be completely electrically isolated from the input, depending on the type of converter that is used. The overall efficiency is not greatly affected by the difference between input and output voltage, and can exceed 90%, minimizing waste heat and enabling the unit to be extremely compact.

A DC-DC converter is an integrated circuit package that includes a high-speed switching device (almost always, a *MOSFET*) in conjunction with an oscillator circuit, an inductor, and a diode. By comparison, a **linear regulator** is usually based around bipolar transistors. Its input must always be at a higher voltage than its output, and its efficiency will be inversely proportional with the voltage drop that it imposes. See the **voltage regulator** entry in this encyclopedia for additional information.

There is no single symbol to represent a DC-DC converter. Some simplified schematics showing the principles of operation of commonly used converters are referenced under the following Variants section.

A DC-DC converter is also typically found in the output stage of a switching **AC-DC power supply**.

How It Works

An internal oscillator controls a *MOSFET* semiconductor that switches the DC input on and off at a high frequency, usually from 50KHz to 1MHz. Output voltage is adjusted by varying the *duty cycle* of the oscillator—the length of each “on” pulse relative to each “off” interval. This is known as *pulse-width modulation*, or *PWM*. The duty cycle is controlled by sampling the output of the converter and using a comparator to subtract the output voltage from a reference voltage, to establish an error value. This is passed to another comparator, which subtracts the error voltage from an oscillator ramp signal. If the error increases, the oscillator signal is more heavily clipped, thus changing the effective ratio of on/off pulse

lengths. A simplified schematic of the PWM circuit is shown in Figure 17-1, which omits other components for clarity. The system of subtracting an error voltage from a ramp oscillator voltage to obtain a pulse-width modulated signal is illustrated in Figure 17-2.

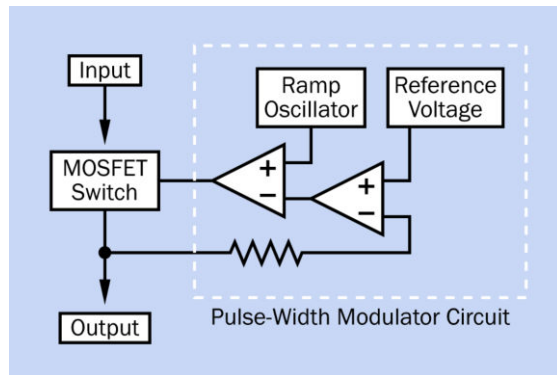


Figure 17-1. The heart of a DC-DC converter is a MOSFET switch, which operates at a high frequency with pulse-width modulation used to create an adjustable DC output.

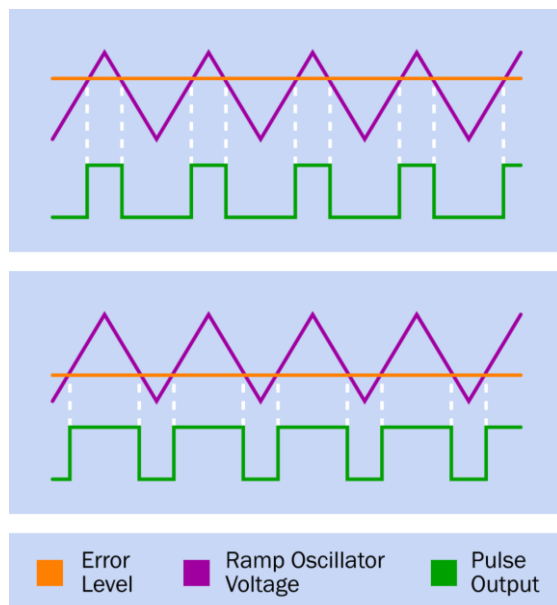


Figure 17-2. To achieve pulse-width modulation, an error-level voltage is established by comparing the output from the converter with a reference voltage. The error level, shown as an orange line, is then subtracted from the output from a ramp oscillator. The pulse width varies accordingly.

The key to the efficiency of a DC-DC converter is an inductor, which stores energy in its magnetic field during “on” pulse and releases it in the discharge phase. Thus, the inductor is used as a temporary reservoir and minimizes the ripple current. All converter variants use a coil for this purpose, although its placement varies in relation to the diode and capacitor that complete the basic circuit.

Variants

Four basic switching circuits are used in DC-DC converters and are defined in the coming sections, with a formula to determine the ratio between input voltage (V_{in}) and output voltage (V_{out}) in each case. In these formulae, variable D is the duty cycle in the pulse train generated through an internal MOSFET switch. The duty cycle is the fraction of the total on-off cycle that is occupied by each “on” pulse. In other words, if T_{on} is the duration of an “on” pulse and T_{off} is the “off” time:

$$D = T_{on} / (T_{on} + T_{off})$$

Buck Converter

See Figure 17-3. The output voltage is lower than the input voltage. The input and output share a common ground. For this circuit:

$$V_{out} = V_{in} * D$$

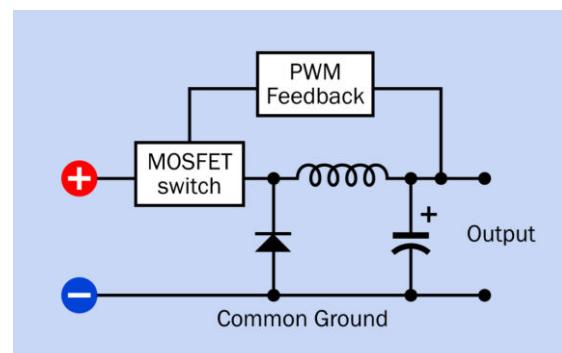


Figure 17-3. Basic topology of a buck-type DC-DC converter.

Boost Converter

See Figure 17-4. The output voltage is greater than the input voltage. The input and output share a common ground. For this circuit:

$$V_{\text{out}} = V_{\text{in}} / (1-D)$$

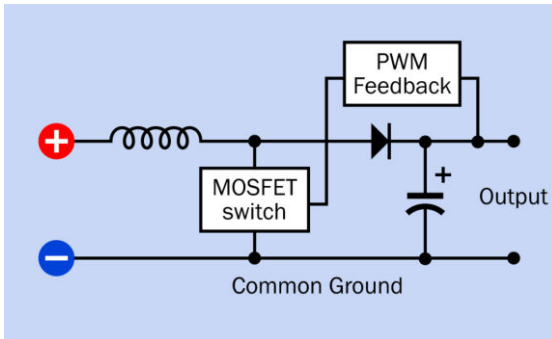


Figure 17-4. Basic topology of a boost-type DC-DC converter.

Flyback Converter with Inductor

Commonly known as a *buck-boost converter*. See Figure 17-5. The output voltage can be less than or greater than the input voltage. The input and output share a common ground. For this circuit:

$$V_{\text{out}} = V_{\text{in}} * (D / (1-D))$$

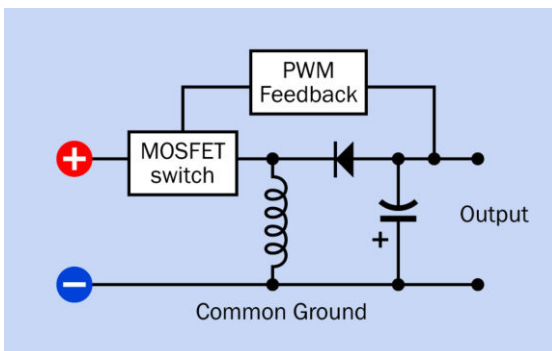


Figure 17-5. Basic topology of a flyback-type DC-DC converter.

Flyback Converter with Transformer

See Figure 17-6. The output voltage can be less than or greater than the input voltage. The input and output are isolated from one another. For this circuit:

$$V_{\text{out}} = V_{\text{in}} * (D / (1-D))$$

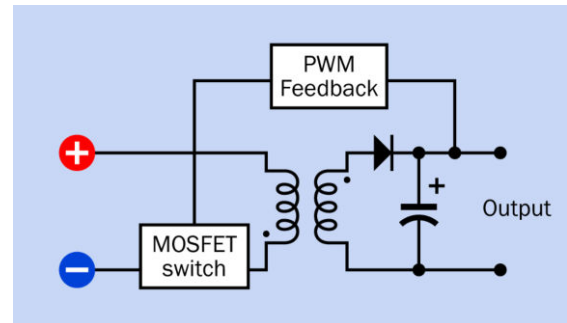


Figure 17-6. Basic topology of a flyback-type DC-DC converter. (Buck, boost, and flyback topologies adapted from Maxim Integrated Products.)

Using a transformer in the converter allows multiple outputs with different voltages, supplied through multiple transformer windings.

Formats

A converter may be packaged in a flat rectangular box that requires no additional heat sink and has pins for through-hole insertion into a PC board. Sizes usually range up to 2" × 2". Power handling can range from 5 to 30 watts. Converters of this type are shown in Figure 17-7. (Top: Input range of 9 to 18VDC, fixed output of 5VDC at 3A completely isolated from the input. Typical efficiency of approximately 80%. The case is made of copper, providing good heat dissipation with electrical shielding. Center: Input range of 9 to 18VDC, fixed output of 5VDC at 500mA completely isolated from the input. Typical efficiency of approximately 75%. The manufacturer claims that external capacitors are only needed in critical applications. Bottom: SIP format, fixed input

of 12VDC, fixed output of 5VDC at 600mA completely isolated from the input. Typical efficiency of approximately 75%. Requires external capacitors for ripple rejection.)



Figure 17-7. A selection of sealed DC-DC converters.

Lower-power converters are also available as surface-mount devices.

Some adjustable-output converters are supplied as multiple surface-mounted components pre-installed on a mini-board that has pins for through-hole insertion in a printed circuit board. Their high efficiency enables them to handle a lot of power for their size. In [Figure 17-8](#), the converter accepts a 4.5 to 14VDC input range and has an adjustable output of 0.6 to 6VDC. It is rated at a surprising 10A or 50W and is more than 90% efficient. However, it draws 80mA in a no-load state, causing it to become quite hot. A thermal cutout or automatic shutdown may be used if the converter will not be driving a consistent load.

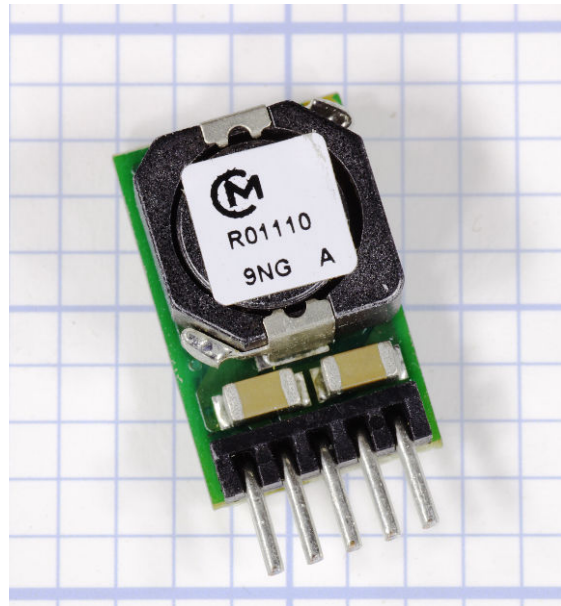


Figure 17-8. An adjustable DC-DC converter rated for 10A or 50W. The output voltage is determined by adding an external resistor or trimmer potentiometer. External smoothing capacitors are required, as shown in the component's datasheet.

The miniboard in [Figure 17-9](#) accepts an input voltage from 7 to 36VDC and has an adjustable output ranging from 2.5 to 12.6VDC, at up to 6A. It is non-isolated (has a common negative bus) and claims to be more than 95% efficient at full load.

The miniboard in [Figure 17-10](#) accepts an input voltage from 4.5 to 14VDC and has an adjustable output ranging from 0.6 to 6VDC at up to 20A. It is non-isolated (has a common negative bus) and claims to be more than 90% efficient at full load.

Values

Relevant values include:

Nominal Input Voltage and Frequency

A wide range of input voltages is often acceptable, as the PWM can vary accordingly. Converters

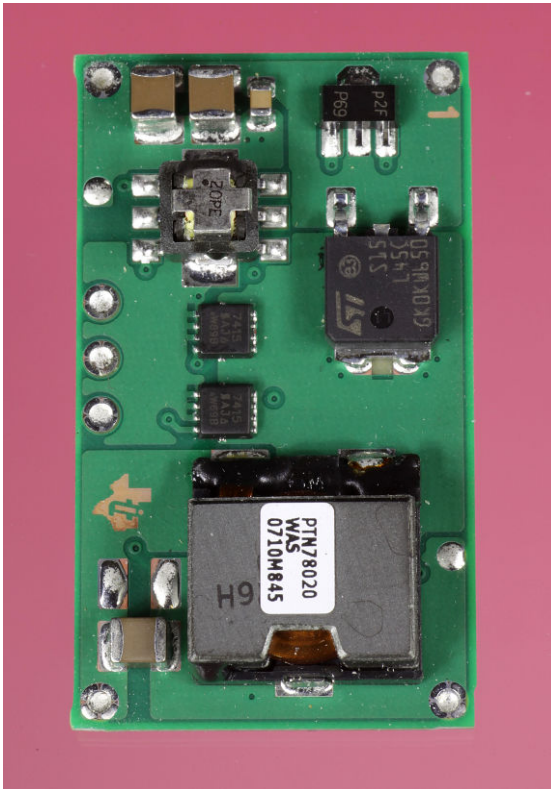


Figure 17-9. Another adjustable DC-DC converter. The output voltage is determined by adding an external resistor or trimmer potentiometer. External smoothing capacitors are required, as shown in the component's datasheet.

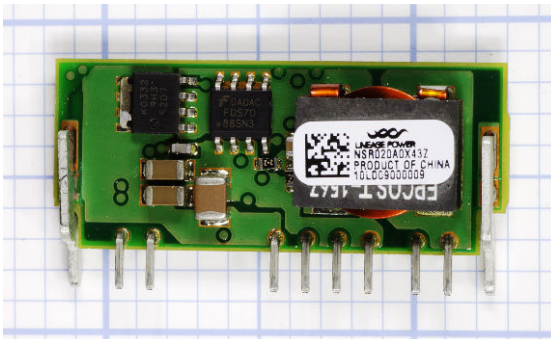


Figure 17-10. Another adjustable DC-DC converter. The output voltage is determined by adding an external resistor or trimmer potentiometer. External smoothing capacitors are required, as shown in the component's datasheet.

often allow equipment to be usable internationally, on any voltage ranging from 100VAC to 250VAC, at a frequency of 50Hz or 60Hz, without any adaptation.

Output Voltage

As previously noted, many converters allow the output voltage to be adjusted by adding an external resistor or potentiometer. Alternatively, there may be multiple fixed output voltages, accessible via different pins on the package. They may also provide a positive voltage and equally opposite negative voltage relative to a ground pin.

Input Current and Output Current

Because input voltage and output voltage are likely to be different, the current alone is not a reliable guide to power handling.

A datasheet should specify input current with no load (open circuit on the output side). This current will have to be entirely dissipated as heat.

Load Regulation

This is usually expressed as a percentage and suggests the extent to which output voltage may be pulled down when the load on a DC-DC converter increases. If V_{nil} is the measured output voltage with no load, and V_{max} is the measured output voltage with the maximum rated load:

$$\text{Load regulation} = 100 * (V_{nil} - V_{max}) / V_{max}$$

However, note that some converters are designed with the expectation that they will never be used with zero load across the output. In these cases, V_{nil} will be the voltage at minimum rated load.

Efficiency

This is a measure of how much input current must be dissipated as heat. A converter with a 12-volt input, drawing a maximum 300mA input current, will consume 3.6 watts (3,600mW). If it is 80% efficient, it will have to dissipate roughly 20% of its power as heat, or 720mW.

Ripple and Noise

Sometimes abbreviated R/N, this may be measured in mV or as a percentage. Check the specification carefully to determine whether the ripple-and-noise values require use of external smoothing capacitors. Often, this is the case.

Isolated or Non-Isolated

This crucial piece of information is often found near the top of a datasheet, not in the detailed specifications.

How to Use it

Because a converter creates electrical noise, it should be prevented from affecting other components by adding substantial *bypass capacitors* as close as possible to its input and output pins. For most converters, external capacitors are mandatory, and their *effective series resistance* (ESR) should be as low as possible (see the **capacitor** entry in this encyclopedia for an explanation of ESR). Tantalum capacitors are preferable to electrolytics for this reason, and are also more durable. Some manufacturers recommend placing a tantalum capacitor in parallel with an electrolytic. A small ceramic capacitor, typically 0.1µF, is often recommended in an addition to larger-value capacitors on the output side.

The voltage rating of each capacitor should be twice the voltage at the point in the circuit where it is used. The capacitance value will usually be higher for higher-current converters. Values of 100µF are common, but for high amperage, a value may be as high as 1,000µF.

While datasheets are often inadequate for some types of components, datasheets for DC-DC converters usually include detailed instructions regarding bypass capacitors. Following these instructions is essential. In the relatively rare instances that a datasheet makes no mention of bypass capacitors for a converter, this does not necessarily mean that the capacitors are unnecessary. The manufacturer may assume that they will be used as a matter of course.

Converters are used in a very wide range of devices, supplying power ranging from a few milliamps to tens of amps. At the lower end of the scale, devices such as cellular telephones, portable computers, and tablets contain subcircuits that require different voltages, some of which may be higher than the voltage of a single battery or battery pack that powers the device. A converter can satisfy this requirement. Because a converter can be designed to maintain a fixed output in response to a range of input voltages, it can also compensate for the gradual decline in voltage that occurs during battery usage.

A boost-type converter can be used to double the voltage from a single 1.5V battery in an LED flashlight where 3 volts are required to power the LED. Similarly, a boost-type converter can provide the necessary voltage to run a cold-cathode fluorescent tube that provides backlighting in an LCD computer display.

On a circuit board that is primarily populated with 5VDC components and is fed by a single 5VDC power supply, a converter can be used to supply 12VDC for one special purpose, such as an analog-digital converter or a serial data connection.

If electromechanical **relays** or other inductive loads share a common ground with components, such as *logic chips* or **microcontrollers**, it may be difficult to protect the sensitive components from voltage spikes. A flyback converter with a transformer separating the output from the input can allow the “noisy” section of the circuit to be segregated, so long as the converter itself does not introduce noise. Since the electromagnetic interference (EMI) introduced by converters varies widely from one model to another, specifications should be checked carefully.

Very low-power components can pick up EMI from the wires or traces leading into and out of a converter. In this type of circuit, adequate noise suppression may be impossible, and a converter may not be appropriate.

What Can Go Wrong

Electrical Noise in Output

Electrolytic capacitors may be inadequate to smooth the high frequencies used. Multilayer ceramic capacitors or tantalum capacitors may be necessary. Check the manufacturer's datasheet for minimum and maximum values. Also check the datasheet for advice regarding placement of capacitors on the input side as well as across the output.

Excess Heat with No Load

Some converters generate substantial heat while they are powered without a load. The manufacturer's datasheet may not discuss this potential problem very prominently or in any detail. Check the input rating, usually expressed in mA, specified for a no-load condition. All of this current will

have to be dissipated as heat, and the very small size of many converters can result in high localized temperatures, especially since many of them allow no provision for a heat sink.

Inaccurate Voltage Output with Low Load

Some converters are designed to operate with at least 10% of full rated load across their output at all times. Below this threshold, output voltage can be grossly inaccurate. Read datasheets carefully for statements such as this: "Lower than 10% loading will result in an increase in output voltage, which may rise to typically double the specified output voltage if the output load falls to less than 5%." Always use a meter to verify the output voltage from a converter at a variety of different loads, and perform this test before installing the converter in a circuit.

DC-AC inverter

A power inverter must not be confused with a *logic inverter*, which functions as a digital component in *logic circuits* to invert the state of a low-voltage DC input from high to low or low to high. Logic inverters are discussed in Volume 2.

OTHER RELATED COMPONENTS

- **AC-DC power supply** (See [Chapter 16](#))
- **DC-DC converter** (See [Chapter 17](#))

What It Does

A power inverter is included here as counterpoint to a **power supply** or *AC adapter*, since it has the opposite function. The inverter receives an input of *direct current* (typically 12VDC from a car battery) and delivers an output of *alternating current* (AC) in the range 110VAC-120VAC or 220VAC-240VAC, suitable to power many low-wattage appliances and devices. The interior of a low-cost inverter is shown in [Figure 18-1](#).

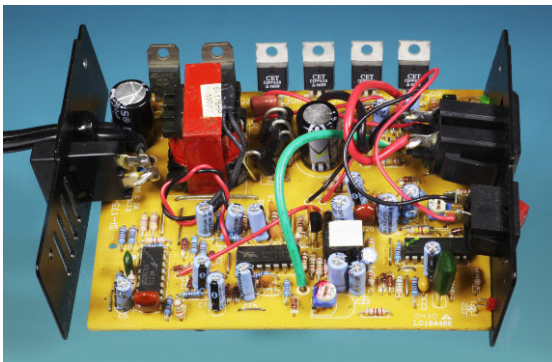


Figure 18-1. The interior components in a 175-watt inverter.

How It Works

The first stage of an inverter typically raises a 12VDC input to a higher DC voltage via an internal **DC-DC converter**, then uses a switching circuit to create an approximation of the sinusoidal profile that is characteristic of AC voltage.

Digital switching components naturally tend to create square waves, whose simple appearance conceals the presence of higher frequencies, or *harmonics*, that are ignored by some devices (especially those that convert electricity into heat) but can be troublesome in consumer electronics equipment. A primary objective of inverter design is to adapt or combine square waves to emulate a classic AC sine wave with reasonable fidelity. Generally speaking, the more accurately an inverter emulates a sine wave, the more expensive it tends to be.

The most primitive inverter would create a plain square wave such as that shown in red in [Figure 18-2](#), superimposed on a comparable sine wave (in green). Note that alternating current rated at 115 volts actually peaks at around 163 volts because the number 115 is the approximate *root mean square (RMS)* of all the voltage values during a single positive cycle. In other words, if the voltage is sampled x times during a cycle, an

RMS value can be derived by squaring each sample, adding all the samples, dividing by x , and then taking the square root of the result. The RMS value is important as a means to calculate actual power delivered because it can be multiplied by the current to obtain an approximate value in watts.

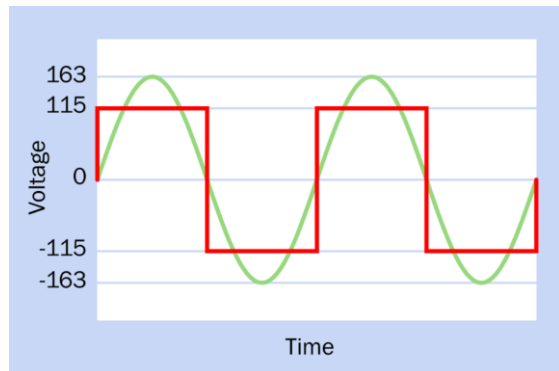


Figure 18-2. Comparison of an AC voltage sine wave (green) and a square wave of the same frequency (red), both delivering a roughly similar amount of power.

Variants

As a first step toward a better approximation of a sine wave, gaps of zero voltage can be inserted between square-wave pulses. This “gapped” square wave is shown in [Figure 18-3](#).

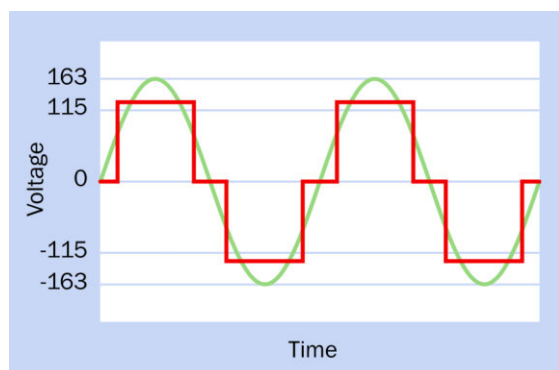


Figure 18-3. Introducing pauses or gaps of zero voltage between square-wave pulses can produce slightly improved resemblance to a sine wave.

A further improvement can be achieved if an additional, shorter pulse of higher voltage is added to each primary pulse, as shown in [Figure 18-4](#). Outputs of this kind are referred to as *modified sine wave*, although they are actually square waves modified to emulate a sine wave. Their inaccuracy is expressed as *total harmonic distortion (THD)*. Some authorities estimate that the THD of gapped square-wave output is around 25%, whereas the addition of shorter square waves reduces this to around 6.5%. This is a topic on which few people agree, but there is no doubt that a “stacked” sequence of square waves provides a closer emulation of a sine wave.

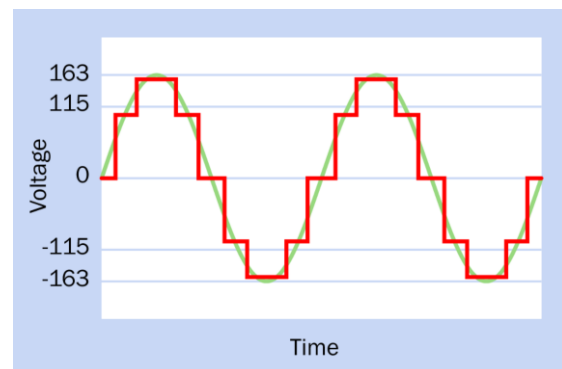


Figure 18-4. A secondary stream of narrower square-wave pulses can improve the fidelity of an inverter's output.

A *true sinewave inverter* typically uses *pulse-width modulation (PWM)* to achieve THD of less than 1%. It generates a stream of pulses much higher in frequency than that of the AC output, and varies their widths in such a way that their averaged voltage closely approximates the voltage variations in a sine wave. A simplified representation of this principle is shown in [Figure 18-5](#).

Values

Small inverters are typically rated to deliver up to 100 watts and may be fitted with a 12VDC plug for insertion in a vehicle's cigarette lighter. Since a cheap inverter may be only 80% efficient, 100 watts at 135VAC will entail drawing as much as

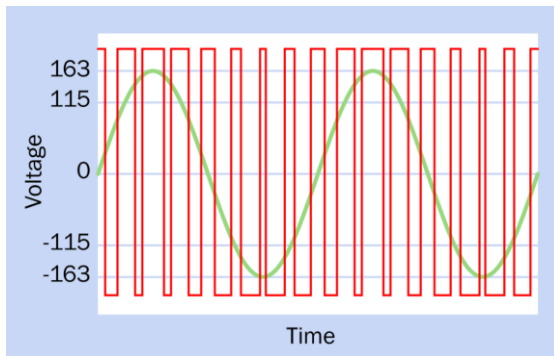


Figure 18-5. Pulse-width modulation adjusts the widths of pulses delivered at a high frequency. The pulse widths can be averaged to generate voltage that follows a close approximation of a sine wave.

10 amps at 12VDC. Cigarette lighters are usually fused at 15 or 20 amps, so 100 watts is a reasonable value. Inverters that are rated above 150 watts usually have cables terminating in oversize alligator clips for direct connection to the terminals of a 12V battery.

While the *cold cranking* rating of a car battery may be 100 amps or more, the battery is only designed to deliver that power for up to 30 seconds at a time. Inverters rated for as much as 500 watts will exceed the normal capacity of a single car battery, although if the battery is mounted in a vehicle, it can be supplemented by running the engine so that the alternator shares some of the load. A 500-watt inverter is better supplied by two or more 12-volt car batteries wired in parallel.

How to Use it

Small inverters are typically used in vehicles to run cellphone chargers, music players, or laptop computers. Large inverters are an integral part of off-the-grid solar and wind-powered systems, where battery power must be converted to AC house current. Uninterruptible power supplies contain batteries and inverters capable of run-

ning computer equipment for a brief period. Battery-driven electric vehicles with AC motors use inverters with an exceptionally high current rating.

There is a lack of consensus regarding possible harmful effects of powering electronics equipment with a low-cost modified sine wave inverter. Anecdotal evidence suggests that where the equipment uses its own *switching power supply* or uses an *AC adapter* (either mounted internally or as an external plug-in package), the filtering built into the power supply will block harmonics from the inverter.

Other evidence suggests that cheap inverters may have adverse effects on devices containing synchronous motors that run direct from AC. There are reports that fluorescent lighting and photographic electronic flash systems may be unsuitable for use even with modified sine wave inverters. However, differences in product design and component quality make it impossible to generalize. A cheaply made inverter may generate a wave form that is not even a close approximation of a square wave. See Figure 18-6.

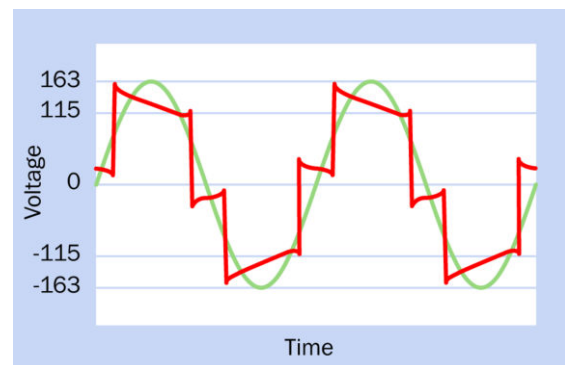


Figure 18-6. A cheaply made inverter can generate a distorted wave form that is even higher in noise than a pure square wave. This sample is adapted from an actual oscilloscope trace.

What Can Go Wrong

If multiple batteries are connected in parallel, using suitably heavy-gauge wire to power a large inverter, the batteries must be identical in specification and age, and must all be equally charged to prevent high and potentially dangerous flows of current among the batteries as they attempt to reach an equilibrium among themselves. Interconnections must be firmly clamped to clean battery terminals. For additional information, see the **battery** entry in this encyclopedia.

Problems associated with inverters are likely to be mundane. A 12V wiring to the inverter can

overheat if items such as clothes or bedding are left on top of it; a high-wattage fan-cooled inverter can overheat if the fan is obstructed by poor placement or impaired by accumulated dirt; alligator clips may become dislodged from battery terminals; and power surges drawn by inductive loads such as motors may trigger the inverter's *breaker*, especially if they are used in conjunction with other equipment.

As always, high amperage should be treated with caution and respect, regardless of it being delivered at "only 12 volts."

voltage regulator

Correctly known as a *linear voltage regulator* to distinguish it from a *switching regulator* or **DC-DC converter**. However, the full term is not generally used, and “voltage regulator” is normally understood to mean a linear voltage regulator.

OTHER RELATED COMPONENTS

- **DC-DC converter** (See [Chapter 17](#))
- **AC-DC power supply** (See [Chapter 16](#))

What It Does

A linear voltage regulator provides a tightly controlled DC output, which it derives from an unregulated or poorly regulated DC input. The DC output remains constant regardless of the load on the regulator (within specified limits). It is a cheap, simple, and extremely robust component.

There is no single schematic symbol for a linear voltage regulator.

The general physical appearance of a commonly used type of regulator, rated for an output of around 1A DC, is shown at [Figure 19-1](#). The LM7805, LM7806, LM7812, and similar regulators in the LM78xx series are encapsulated in this type of package, with pins that are spaced at 0.1” and have functions as shown. Other types of regulator may differ in appearance, or may look identical to this one but have different pin functions. Always check datasheets to be sure.

How It Works

All linear regulators function by taking some feedback from the output, deriving an error value by comparing the output with a reference voltage (most simply provided by a zener diode),

and using the error value to control the base of a *pass transistor* that is placed between the input and the output of the regulator. Because the transistor operates below saturation level, its output current varies linearly with the current applied to its base, and this behavior gives the linear regulator its name. [Figure 19-2](#) shows the relationship of these functions in simplified form; [Figure 19-3](#) shows a little more detail, with a *Darlington pair* being used as the pass transistor. The base of the pair is controlled by two other transistors and a comparator that delivers the error voltage. This version of a voltage regulator is known as the *standard type*.

The voltage difference required between the base and emitter of an NPN transistor is a minimum of 0.6V. Because multiple transistors are used inside a standard-type voltage regulator, it requires a minimum total voltage difference, between its input and its output, of 2VDC. This voltage difference is known as the *dropout voltage*. If the voltage difference falls below this minimum, the regulator ceases to deliver a reliable output voltage until the input voltage rises again. *Low dropout regulators* allow a lower voltage difference, but are more expensive and less commonly used. They are described under the following Variants section.

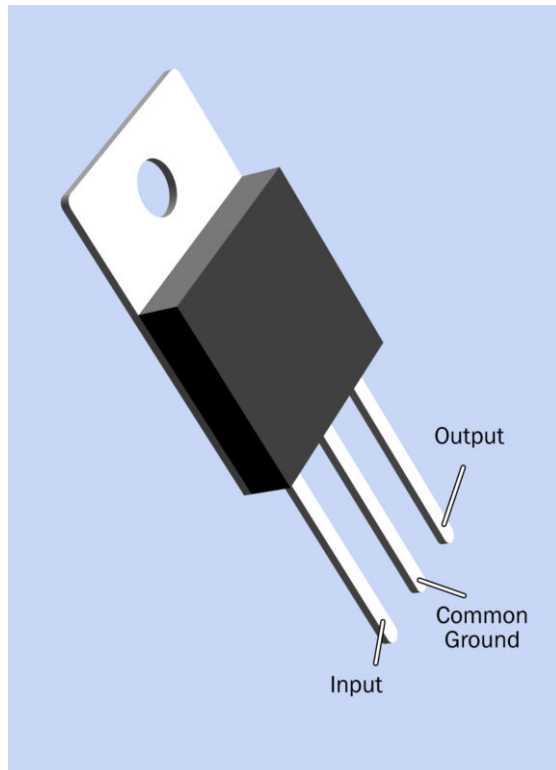


Figure 19-1. The package design of a commonly used voltage regulator. Others may be significantly different, and the pin functions may vary. Check manufacturer datasheets for details.

Discrete components could in theory be used to build a voltage regulator, but this ceased to be cost-effective several decades ago. The term is now understood to mean one small integrated package containing the basic circuit augmented with additional, desirable features, such as automatic protection against overload and excessive heat. Instead of burning out if it is overloaded, the component simply shuts down. Most voltage regulators also tolerate accidentally reversed power connection (as when batteries are inserted the wrong way around) and accidentally reversed insertion of the regulator in a circuit board.

Other components can satisfy the requirement to deliver power at a reduced voltage. Most simply, if two resistors in series are placed across a

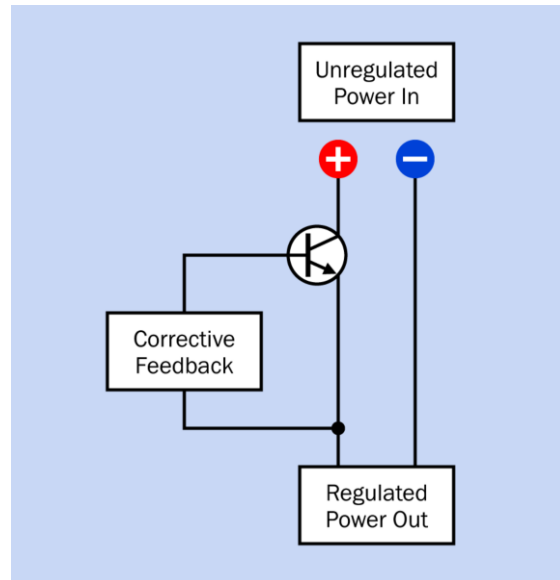


Figure 19-2. A linear voltage regulator basically consists of a transistor whose base is controlled by corrective feedback derived from the output.

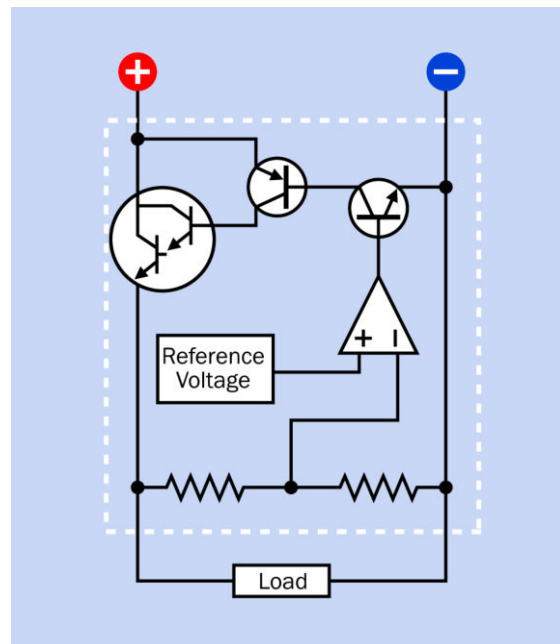


Figure 19-3. The fundamental internal features of a standard-type voltage regulator, including a Darlington pair, two transistors, a voltage divider, comparator, and reference voltage source, shown inside the dashed white line.

power source, they form a **voltage divider**, which provides an intermediate voltage at the connection between them. However, this voltage will vary depending on fluctuations in the input voltage and/or load impedance. A voltage regulator is the simplest way to supply a voltage that remains stable regardless of excursions in the input or fluctuations in power consumed by the load.

The disadvantage of a standard-type voltage regulator is that it is inefficient, especially when a relatively high input voltage is used to deliver a relatively low output voltage. If V_{in} is the input voltage, V_{out} is the output voltage, and I_{out} is the output current, the average power loss, P , is given by the formula:

$$P = I_{out} * (V_{in} - V_{out})$$

For example, if the output current is 1A, the input voltage is 9VDC, and the output is 5VDC, 44% of the input power will be wasted, and the component will be only 56% efficient. The wasted power (about 4 watts, in this case) will be dissipated as heat. Even when a standard-type regulator runs at its minimum 2VDC dropout voltage, it must dissipate 1W when delivering 0.5A.

Variants

Packaging

The package for the LM78xx series of regulators, shown in [Figure 19-1](#), incorporates an aluminum plate drilled with a hole so that it can be bolted to a heat sink. Voltage regulators with a lower rated maximum output current (typically, 100mA) do not have the same need for a heat sink, and are available in a package that resembles a small transistor.

Some integrated circuits are available containing two voltage regulators, electrically isolated from each other.

Popular Varieties

In the LM78xx series, the last two digits in the part number specify the output voltage, which is fixed. Thus the LM7805 delivers 5VDC, the

LM7806 delivers 6VDC, and so on. For regulators with a fractional voltage output (3.3VDC being common), an additional letter may be inserted in the part number, as in the 78M33.

Many copies of the LM78xx series are made by different manufacturers, the copies being functionally identical, regardless of additional letters that are added to the part number to identify its source or other attributes.

The LM78xx regulators are mostly rated to be accurate within 4%, although actual samples almost always deliver voltages that are more precise than this range suggests.

Adjustable Regulators

While the majority of regulators have a fixed output, some allow the user to set the output by adding one or more resistors. The LM317 is a popular example. Its output voltage can range from 1.25VDC to 37VDC and is set via a resistor and a trimmer potentiometer, as illustrated in [Figure 19-4](#). If $R1$ is the fixed-value resistor and $R2$ is the trimmer, as shown in the schematic, the output voltage, V_{out} , is given by the formula

$$V_{out} = 1.25 * (1 + (R2 / R1))$$

Typical values for $R1$ and $R2$ would be 240Ω and 5K, respectively. With the trimmer at the middle of its range, V_{out} would be $1.25 * (1 + (2500 / 240))$ = approximately 15VDC, requiring an input of at least 17VDC. However, if the trimmer is reduced to 720Ω, the output would be 5VDC. In practice, the value of a trimmer should be chosen so that a mid-range setting provides approximately the desired output. This will enable fine adjustment of the output voltage.

While the versatility of an adjustable regulator is desirable, its overall power dissipation is still proportional to the difference between the input voltage and the output voltage. To minimize heat loss, this difference should not exceed the dropout voltage by a larger amount than is absolutely necessary.

An adjustable regulator may require larger bypass capacitors than a regulator with a fixed output. A manufacturer's recommendations for the LM317 are shown in [Figure 19-4](#).

Negative and Positive Regulators

While most linear voltage regulators are designed for "positive input" (conventional current flow from input to output), some are intended for "negative input." In this variant, the common terminal is positive, and the input and output are negative in relation to it.

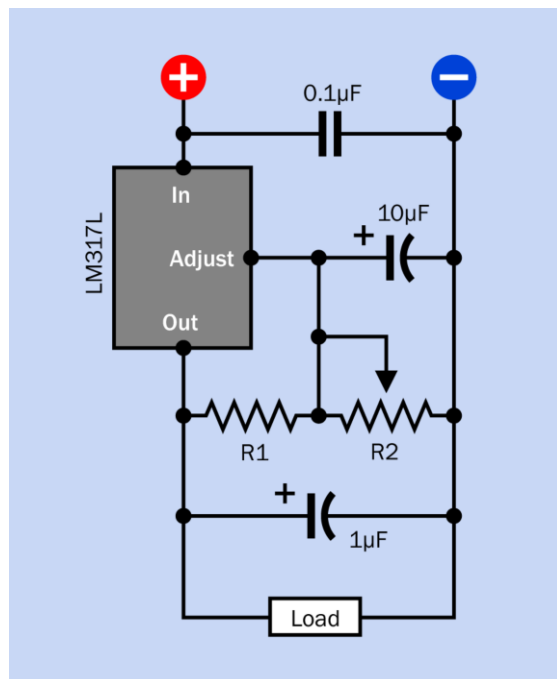


Figure 19-4. Schematic for the LM317L adjustable voltage regulator, based on a circuit recommended by National Semiconductor, with bypass capacitors added for ripple rejection.

Low-Dropout Linear Regulators

Low-dropout regulators (sometimes referred to as *LDO* regulators) allow a much lower dropout voltage by using a single PNP or MOSFET transistor. LDO regulators are popularly used in battery-powered devices where efficiency should be maximized and heat dissipation should be minimized. For example, the LM330 is

a regulator with a 5VDC output, tolerating a dropout voltage of 0.6V, allowing it to be used with four AAA cells. In an LDO regulator the dropout voltage actually varies with load current and may diminish to as little as one-tenth of its rated value when the output current is minimal.

The majority of low-dropout regulators are sold in surface-mount packages, and are designed for maximum output of 100mA to 500mA. Only a few exceptions exist. They tend to be slightly more expensive than regulators with the typical 2V dropout rating.

Three voltage regulators are shown in [Figure 19-5](#). From left to right, they are rated 5VDC at 1A, 12VDC at 1A, and 5VDC at 7.5A. The two smaller regulators are of the LM78xx series. The larger regulator claims a low maximum dropout voltage of 1.5VDC, and its output voltage can be adjusted with an external potentiometer and resistor.

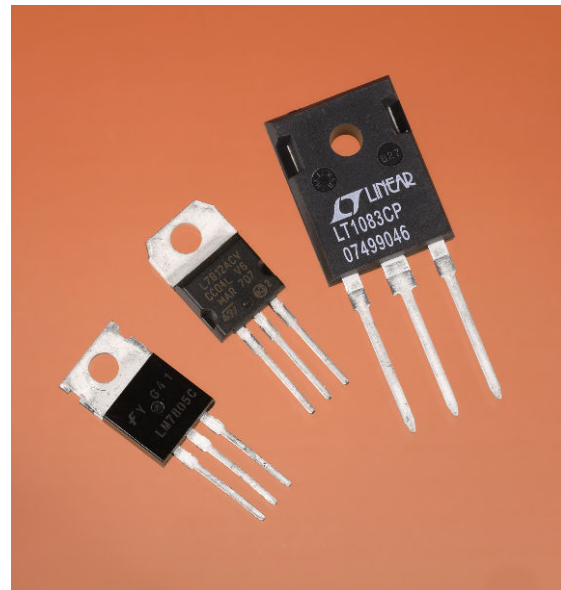


Figure 19-5. Two voltage regulators from the LM78xx series, and a third high-current, low-dropout, adjustable regulator rated 5VDC (adjustable upward) at 7.5A.

Quasi-Low-Dropout Linear Regulators

Where a standard regulator uses a Darlington pair as the pass transistor and an LDO uses a single PNP transistor, the so-called Quasi-LDO uses a combination of NPN and PNP transistors and has an intermediate dropout voltage, typically a maximum of 1.5VDC. However, the terms LDO and Quasi-LDO are not used uniformly in the industry. One manufacturer markets Quasi-LDO regulators as LDO regulators, and describes its LDO regulators as *Very Low Dropout* regulators. Consult datasheets to determine the actual specification of the product, regardless of its classification.

Additional Pin Functions

Some voltage regulators include an extra pin, typically known as an *enable pin*, which switches off the device in response to a signal from a microcontroller or logic gate.

Some regulators offer another option, an additional *status pin* that can signal a microcontroller that an error mode exists if the regulator output falls significantly below its rated value.

In battery-powered devices, a low-battery sensor is a desirable feature, since a regulator may simply shut down without warning if the input voltage is insufficient. A few regulators, such as the LP2953, provide a low-battery warning output via an extra pin.

Values

Linear voltage regulators with a single, fixed output are commonly available to supply DC outputs of 3.3, 5, 6, 8, 9, 10, 12, 15, 18, and 24 volts, with a few variants offering fractional values in between. The most commonly used values are 5, 6, 9, 12, and 15 volts. The input voltage may be as high as 35VDC.

Maximum output current is typically 1A or 1.5A, in the traditional three-pin, through-hole, TO-220 format. A surface-mount version is available. Other surface-mount formats have lower power limits.

Accuracy may be expressed as a percentage or as a figure for load regulation in mV. A typical load regulation value would be 50mV, while voltage regulation accuracy ranges from 1% to 4%, depending on the manufacturer and the component. While low-dropout regulators are generally more efficient, they do require more ground-pin current. This is not usually a significant factor.

How to Use it

Some components, such as many old-design CMOS chips or the traditional TTL version of the 555 timer, allow a wide range of acceptable input voltages, but most modern logic chips and microcontrollers must have a properly controlled power supply. Regulators such as the LM7805 are traditionally used to provide this, especially in small and relatively simple devices that draw a moderate amount of current, have a low component count, and are powered via a battery or an **AC adapter**. A fully fledged switching power supply is overkill in this kind of application.

A linear voltage regulator cannot respond instantly to changes in input voltage. Therefore, if the input supply contains voltage spikes, these spikes may pass through the regulator. Bypass capacitors should be applied preventively. A sample schematic showing an LM7805 regulator with bypass capacitors recommended by a manufacturer is shown in [Figure 19-6](#).

In a battery-powered device where standby power is required for long periods and full power is only needed intermittently, the *quiescent current* drawn by a minimally loaded voltage regulator is important. Modern LDO regulators may draw as little as 100µA when they are very lightly loaded. Other types may consume significantly more. Check datasheets to find the most appropriate component for a particular application.

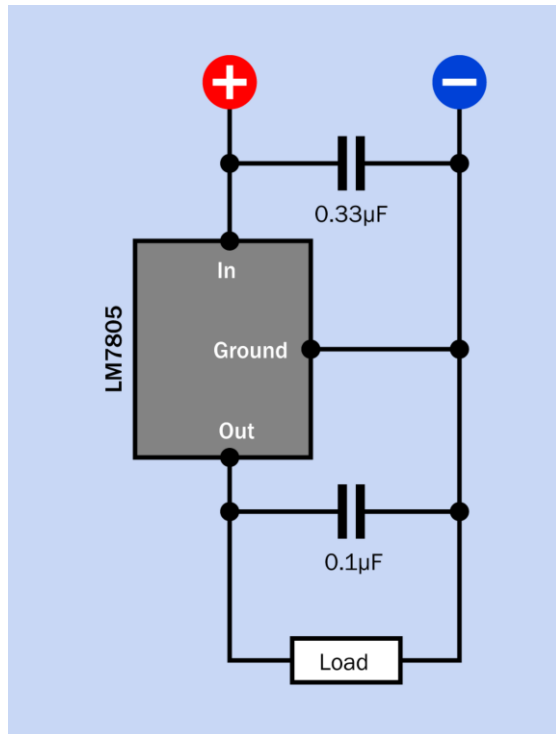


Figure 19-6. Typical schematic for use of an LM7805 regulator, with capacitor values based on recommendations from Fairchild Semiconductor.

Note that DC-DC power **converters** may draw a lot of current when they are lightly loaded, and will dissipate large amounts of heat as a result. An LDO is therefore preferable in this situation.

What Can Go Wrong

Inadequate Heat Management

The ability to “dial up” a wide range of voltages from an adjustable regulator such as the LM317 can be a temptation to use it on a “one size fits all” basis, to deliver any output ranging from 5VDC to 18VDC from a uniform 24VDC input. Assuming 1A output current, the worst-case power dissipation in this scenario would be almost 20W. To achieve reasonable efficiency and maintain waste heat at a manageable level, the input voltage should not exceed the output voltage by much more than the dropout voltage.

Even when a voltage regulator is used correctly, it can generate more heat than was expected if the requirements of a circuit are altered during development. An initial handful of components may draw only 100mA, but as more capabilities are requested and more parts are added (especially relays or LED displays) the power consumption can quickly add up, generating an unexpected amount of waste heat and raising the possibility of a sudden (and mysterious) shutdown if the regulator does not have an adequate heat sink.

Transient Response

When there is a major fluctuation in the demand by the load (for example, if an inductive device is switched on elsewhere in the circuit), the voltage regulator requires a finite time to adjust itself and maintain its specified output voltage. This time lag is known as its *transient response*. If a momentary fluctuation is likely, and other components may be sensitive to it, a larger capacitor should be used between the output of the voltage regulator and ground.

The transient response time may also be insufficient to block sudden, brief spikes in input voltage. This may occur, for example, when a low-cost AC adapter that does not have a properly smoothed output is used as the power source. Additional 1µF bypass capacitors may be added at the input and output of a regulator to provide better protection from power fluctuations.

Misidentified Parts

Many types of linear voltage regulators appear physically identical. Care is needed to distinguish those which have fixed output from those that allow a variable output. When using the LM78xx series, double-check the last pair of digits in the part number, which provide the only guide regarding the output. Using an LM7808 instead of an LM7805 may be sufficient to destroy all the 5VDC chips in a logic circuit. It is advisable to use a meter to check the output of any power supply before connecting it with a circuit.

Misidentified Pins

The LM78xx series of voltage regulators uses a very intuitively obvious and consistent scheme for the functions of its pins: input on the left, ground in the center, and output on the right, when looking at the regulator from the front, with its pins facing downward. Unfortunately the consistency of this scheme can encourage an unthinking habit for making connections. The LM79xx series of negative voltage regulators swaps the identity of the input and ground pins, whereas adjustable regulators use yet another different scheme. Good practice suggests checking a component against the manufacturer's datasheet before connecting it.

Dropout Caused by Low Battery

If a regulator rated to deliver 6VDC has a 2VDC dropout voltage and is powered from a 9V battery, the battery can easily drop below the minimum acceptable 8VDC if it becomes old or depleted. When this happens, the output from the regulator will tend to fall, or may oscillate.

Inaccurate Delivered Voltage

A voltage regulator maintains its output voltage between its output pin and ground pin. Thin traces on a circuit board, or a long run of very small-gauge wiring, can impose some electrical resistance, reducing the actual voltage delivered to a component. Ohm's Law tells us that the voltage drop imposed by a trace (or thin wire) will be proportional to the current flowing through it. For example, if the resistance between the output pin of a voltage regulator and a component is 0.5Ω and the current is 0.1A, the voltage drop will be only 0.05V. But if the current increases to 1A, the voltage drop is now 0.5V. Bearing this in mind, a linear voltage regulator should be positioned close to voltage-sensitive components. In printed circuit designs, the traces that deliver power should not have significant resistance.

When using linear voltage regulators with adjustable output, there may be a temptation to

connect adjustment resistor R1 to the positive end of the load, to obtain a "more accurate" delivered voltage. This configuration will not produce the desired result. R1 should always be connected as closely as possible between the output pin and the adjustment pin of the voltage regulator, while R2 should connect between the adjustment pin and the negative end of the load. This is illustrated in Figure 19-7, where the gray wire in each schematic indicates that it possesses significant resistance.

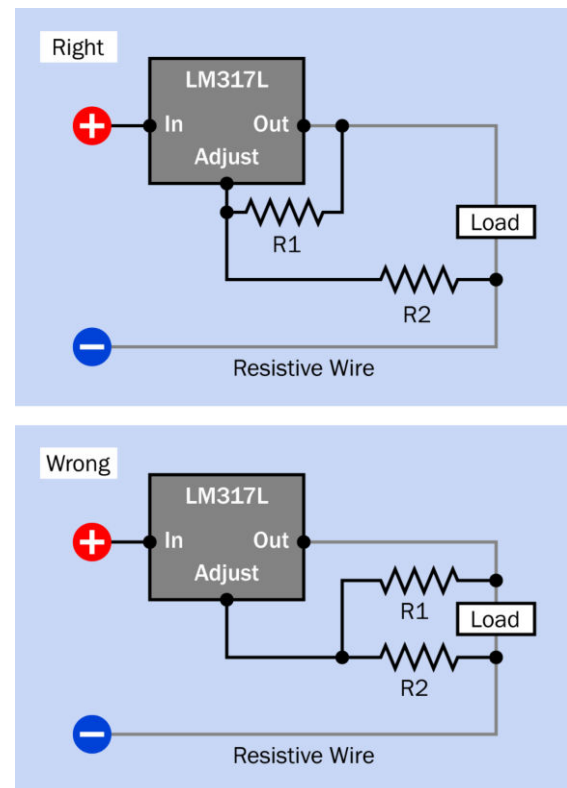


Figure 19-7. When the connection between an adjustable-output voltage regulator and load components has a significant resistance (shown here as a gray "resistive wire"), R1 should always be connected as closely as possible to the pins of the regulator, as shown in the upper schematic. (Derived from schematics prepared by National Semiconductor.)

electromagnet

The term **electromagnet** is used here to mean a coil containing a core of ferromagnetic material that does not move relative to the coil. The core is used solely to create a magnetic field that attracts or repels other parts that have appropriate inherent magnetic properties. Where a center component moves in response to the magnetic force created by current through a coil, this is discussed in the **solenoid** entry. By comparison, the **inductor** entry describes a coil that may or may not have a ferromagnetic core, and is used for the specific purpose of creating reactance, or self-inductance, in an electronic circuit, often in association with alternating current and in combination with resistors and/or capacitors. The **inductor** entry contains a basic discussion and explanation of magnetic force created by electricity.

OTHER RELATED COMPONENTS

- **solenoid** (See [Chapter 21](#))
- **electromagnet** (See [Chapter 20](#))
- **DC motor** (See [Chapter 22](#))
- **AC motor** (See [Chapter 23](#))

What It Does

An electromagnet consists of a coil that creates a magnetic field in response to an electric current. The field is channeled and reinforced by a core of magnetic material (that is, material that can be magnetized). Electromagnets are incorporated in motors, generators, loudspeakers, microphones, and industrial-sized applications such as mag-lev trains. On their own, they provide a means for electric current to hold, lift, or move objects in which a magnetic field can be induced.

A very small, basic electromagnet about 1 inch in diameter is shown in [Figure 20-1](#). No specific schematic symbol for an electromagnet exists, and the symbol for an induction coil with a solid core is often used instead, as shown in [Figure 14-1](#) (the center variant of each of the three) in the **inductor** entry of this encyclopedia.



Figure 20-1. An electromagnet approximately 1 inch in diameter, rated to draw 0.25A at 12VDC.

How It Works

Electric current flowing through a circle of wire (or a series of connected loops that form a helix

or *coil*) will induce a magnetic field through the center. This is illustrated in the **inductor** entry of this encyclopedia, specifically in diagrams [Figure 14-3](#), [Figure 14-4](#), [Figure 14-5](#), and [Figure 14-6](#).

If a stationary piece of ferromagnetic material is placed in the center of the circle or coil, it enhances the magnetic force because the *reluctance* (magnetic resistance) of the material is much lower than the reluctance of air. The combination of the coil and the core is an electromagnet. This is illustrated in [Figure 20-2](#). For a lengthier discussion of this effect, see “[Magnetic Core](#)” (page 122).

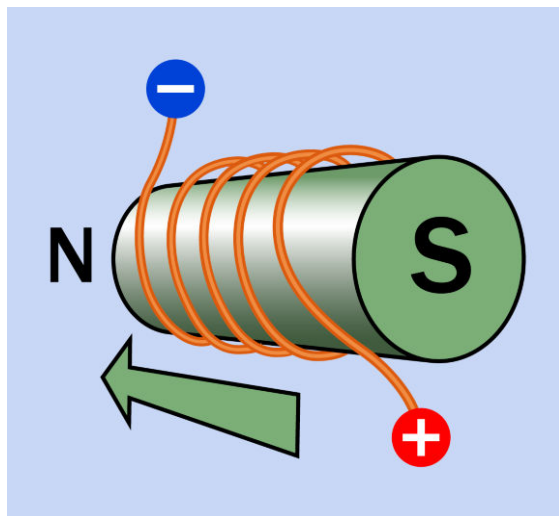


Figure 20-2. Direct conventional current flowing through a wire coiled around a ferromagnetic rod induces a magnetic force in the rod, conventionally considered to flow from south to north.

The magnitude of the electromagnetic *flux density* will be proportional to the current flowing through the coil, assuming a DC power source.

Variants

Electromagnet designs vary according to their application. The simplest design consists of a single coil wound around a rod which may termi-

nate in a plate for applications such as lifting scrap metal. This design is relatively inefficient because the magnetic circuit is completed through air surrounding the electromagnet.

A more efficient, traditional design consists of a U-shaped core around which are wound one or two coils. If the U-shaped core is smoothly curved, it resembles a *horse-shoe magnet*, as shown in [Figure 20-3](#). This design has become relatively uncommon, as it is cheaper to make windings across two separate, straight vertical cores and bridge them. However, the horseshoe configuration is extremely efficient, as the coils induce north and south magnetic polarities in the open ends of the U-shaped core, and the magnetic circuit is completed through any object that is attracted toward the open ends and links them. The attracted object is shown as a rectangular plate in [Figure 20-3](#). Because a magnetic circuit will naturally attempt to limit its extent, and because this goal will be achieved when the circuit is completed, the attractive force of the U-shaped magnet is maximized.

An electromagnet powered by direct current naturally produces a consistently polarized, stable magnetic field. When AC current is applied, an electromagnet may still be used to exert an attractive force on a passive object that is not magnetized but is capable of being magnetized. The electromagnet will change its polarity at almost the same frequency as the AC, and will induce equal and opposite fluctuating polarity in the target, causing mutual attraction. The core of the magnet will be composed of plates separated by thin layers of insulation to inhibit the eddy currents induced by the AC, but still an AC-powered electromagnet will be less efficient than a comparable DC-powered electromagnet because it will also suffer from *hysteresis* as power is consumed by repeatedly reversing the polarity of the *magnetic domains* in the core.

Some electromagnets that are described as suitable for AC power actually contain *rectifiers* that convert the AC to DC.

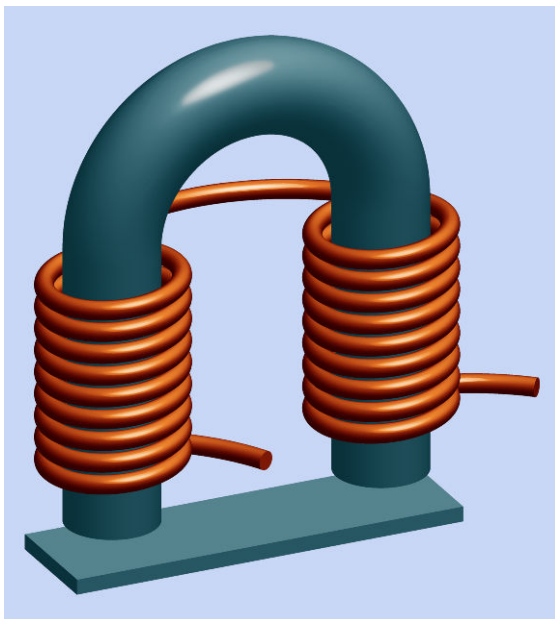


Figure 20-3. This traditional design for an electromagnet has a pedigree stretching back for more than a century. It maximizes efficiency by completing a magnetic circuit through any object that the magnet attracts.

Values

Electromagnets are typically calibrated in terms of their power consumption and retaining force (the weight of an iron target that they can support). The retaining force is usually measured in grams or kilograms.

How to Use it

Electromagnets are used mostly as subassemblies in other components, such as motors and generators, relays, loudspeakers, and disk drives. They have also been used in audio (and video) tape recorders to magnetize ferric oxide on tape, using a magnetic field of varying strength to record an audio signal. In this application, a form of horseshoe magnet with an extremely narrow gap is used, the width of the gap determining the highest frequency that the electromagnet can record, in conjunction with the speed of the tape moving past the head.

The tape recording process can be reversed when the electromagnet “reads” the tape and turns the signal back into a weak alternating current that can be amplified and reproduced through a loudspeaker.

A simple application for an electromagnet is in a traditional-style doorbell, where one or two coils attract a spring-loaded lever, at the top of which is a knob that hits a bell. When the lever is pulled toward the bell, it breaks a contact that supplies power to the electromagnet. This allows the lever to spring back to its original position, which re-establishes the circuit, repeating the process for as long as power is applied to the bell. The bulk and weight of the component parts in this type of doorbell are making it obsolete, as electronic versions containing small loudspeakers become relatively cheaper. However, a **solenoid** may still be used in the type of bell that creates a single chime or pair of chimes.

In any device using a *cathode-ray tube*, electromagnetic coils are used to form a *yoke* around the neck of the tube, to deflect the beam of electrons on its way to the screen. A similar principle is used in electron microscopes. In some cases, electrostatically charged plates are used to achieve the same purpose.

An electromagnet may be used to activate a *reed switch* (the diagram in *Figure 9-7* shows such a switch). In this application, the combination of the electromagnet and the switch are functioning as a **relay**.

When an electromagnet is energized by alternating current, it can be used to *degauss* (in other words, to demagnetize) other objects. The AC is either applied with diminishing current, so that the alternating magnetic polarities gradually subside to zero, or the electromagnet is gradually moved away from the target, again reducing the magnetic influence to (virtually) zero. This latter procedure may be used periodically to demagnetize record and replay heads on tape recorders, which otherwise tend to acquire residual magnetism, inducing background hiss on the tape.

Traditional large-scale applications for electromagnets tend to involve lifting and moving heavy objects or scrap metal, such as junked cars. A more modern application is in magnetic resonance imaging (MRI), which has revolutionized some areas of medicine.

Very large-scale applications for electromagnets include particle accelerators, in which multiple magnetic coils are energized sequentially, and fusion-power generators, where high-temperature plasma is contained by a magnetic field.

What Can Go Wrong

Because an electromagnet requires constant power to maintain its magnetic force, yet it is not doing any actual work so long as its target remains stationary (in contact with the core of the magnet), the current running through the coil of the magnet must be dissipated entirely as heat. Further discussion of this issue will be found at [“Heat” \(page 177\)](#) in the **solenoid** section of this encyclopedia.

solenoid

21

The term **solenoid** was historically used to describe any coil without a magnetic core. More recently and more commonly it describes a coil inside of which a cylindrical plunger moves in response to the magnetic field generated by the coil. In this encyclopedia, the term **electromagnet** has its own entry, and describes a coil with a center component of ferromagnetic material that does not move relative to the coil. It is used solely to attract or repel other parts that have inherent magnetic properties. By comparison, the **inductor** entry describes a coil that is used for the specific purpose of creating reactance, or self-inductance, in an electronic circuit, often in association with alternating current and in combination with resistors and/or capacitors. The **inductor** entry contains a basic discussion and explanation of magnetic force created by electricity.

OTHER RELATED COMPONENTS

- **inductor** (See [Chapter 14](#))
- **solenoid** (See [Chapter 21](#))

What It Does

A typical solenoid consists of a hollow coil inside a *frame*, which may be a sealed cylinder or box-shaped with open sides. In the case of a cylinder, its opposite ends may be referred to as *pole faces*.

At least one of the pole faces has a hole through which a *plunger* (also known as an *armature*) is pulled or pushed by the solenoid. Thus, the solenoid is a device for applying a linear mechanical force in response to current passing through it. In most solenoids, current must be maintained in order to maintain the mechanical force.

A small open-frame solenoid is pictured in [Figure 21-1](#). The upper section of the figure shows the three basic parts: frame, compression spring, and plunger. The lower part of the figure shows the parts assembled.

A larger, closed, cylindrical solenoid is shown in [Figure 21-2](#), with the plunger and spring removed.

A 3D rendering showing a simplified, imaginary, cylindrical solenoid cut in half appears in [Figure 21-3](#). The diagram includes a gray cylindrical shell, often described as the frame; the coil, shown in orange; the plunger, which is pulled into the coil by its magnetic field; and the triangular stop, which limits the plunger's upward travel. The frame of the solenoid exists not merely to protect the coil, but to provide a magnetic circuit, which is completed through the plunger.

The lower end of the plunger is often fitted with a nonmagnetic yoke or perforated plate for connection with other components. Stainless steel can be used for this purpose. The stop may be fitted with a thrust rod (also fabricated from stainless steel) if the solenoid is intended to "push" as well as "pull." Springs to adjust the force of the plunger, or to return it to its initial position when the current through the coil is interrupted, are not shown in the rendering.



Figure 21-1. A small 12VDC solenoid.



Figure 21-2. A larger solenoid rated for 24VDC.

Because there is no standardized schematic symbol for a solenoid, and because this type of component is so widely used in conjunction with valves, any diagram involving solenoids is more likely to emphasize fluid or gas flow with symbols that have been developed for that purpose. In such circuits, a solenoid may be represented simply by a rectangle. However, the symbols shown in [Figure 21-4](#) may occasionally be found.

How It Works

Current flowing through the coil creates a magnetic force. This is explained in the **inductor** entry of this encyclopedia, using diagrams in [Figure 14-3](#), [Figure 14-4](#), [Figure 14-5](#), and [Figure 14-6](#).

If the plunger is fabricated from a material such as soft iron, the coil will induce an equal and opposite magnetic polarity in the plunger. Consequently the plunger will attempt to occupy a position inside the coil where the ends of the plunger are equal distances from the ends of the coil. If a collar is added to the free end of the plunger, this can increase the pulling force on the plunger when it is near the end of its throw because of the additional magnetic pull distributed between the collar and the frame of the solenoid.

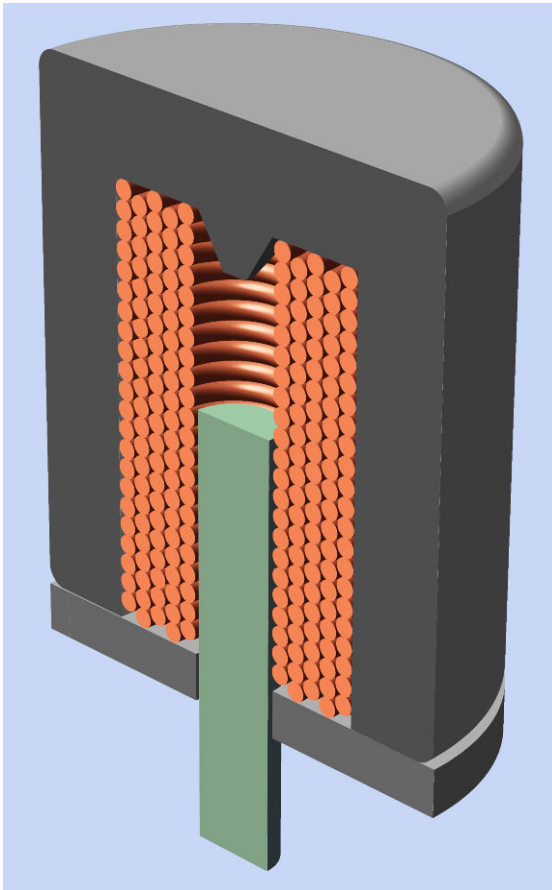


Figure 21-3. A simplified view of a solenoid cut in half, showing the primary parts.

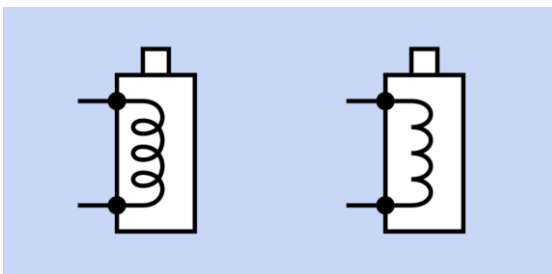


Figure 21-4. Although no standard schematic symbol for a solenoid exists, these symbols may sometimes be found.

A spring can be inserted to apply some resistive force to compensate for the increase in pulling

force that occurs as a larger proportion of the plunger enters the coil. A spring may also be used to eject the plunger, partially at least, when current to the coil is interrupted.

If the plunger is a permanent magnet, reversing DC current to the coil will reverse the action of the plunger.

A solenoid with a nonmagnetized plunger may be energized by AC current, since polarity reversals in the magnetic field generated by the coil will induce equal and opposite reversals in the polarity of the plunger. However, the force curve of an AC-powered solenoid will be different from the force curve of a DC-powered solenoid. See [Figure 21-5](#). The alternating current is likely to induce humming, buzzing, and vibration.

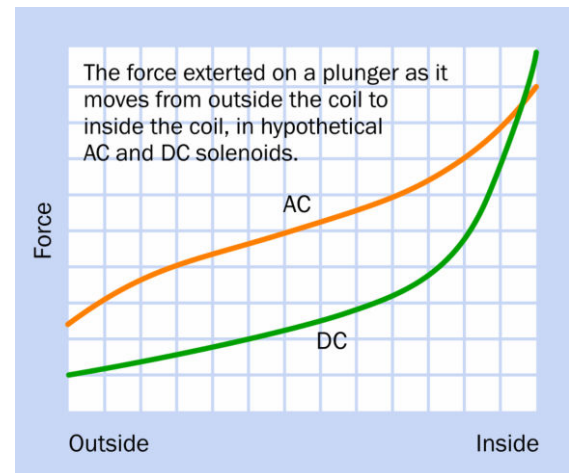


Figure 21-5. A comparison of the force exerted on a plunger, relative to its position as it enters the coil, in hypothetical AC and DC solenoids.

The frame of the solenoid increases the magnetic power that the coil can exert by providing a magnetic circuit of much lower *reluctance* than that of air (reluctance being the magnetic equivalent of electrical resistance). For a lengthier discussion of this effect, see [“Magnetic Core”](#) ([page 122](#)) in the **inductor** entry of this

encyclopedia. If current flowing through the coil increases to the point where the frame becomes magnetically *saturated*, the pulling power of the solenoid will level off abruptly.

The heat generated by a solenoid when it is maintained in its energized state may be reduced if the manufacturer includes a series resistor and a switch that functions as a *bypass switch*. The switch is normally closed, but is opened mechanically when the plunger reaches the end of its throw, thus diverting electricity through the series resistor. This itself will generate some heat as a result of the current flowing through it, but by increasing the total resistance of the system, the total heat output will be reduced. The resistor value is chosen to provide the minimum power needed to retain the plunger at the end of its throw.

Variants

The most common variant is tubular, with open-frame as a secondary option. A tubular solenoid has been shown in [Figure 21-2](#).

Additional variants include:

Low Profile

A shorter, fatter solenoid which may be used if a short throw is acceptable.

Latching

A permanent magnet holds the plunger when it reaches the end of its travel, and continues to hold it after power to the solenoid is disconnected. The plunger itself is also a permanent magnet, and is released by running current of reverse polarity through the coil.

Rotary

This variant is similar in principle to a brushless DC motor and causes the armature to rotate through a fixed angle (typically ranging from 25

to 90 degrees) instead of moving linearly. It is used as a mechanical indicator in control panels, although it is being displaced by purely electronic indicators.

Hinged Clapper

Instead of a plunger, a small hinged panel (the “clapper”) moves in when the solenoid is active, and springs back when the power is interrupted.

Values

The stroke length, duty cycle, and holding force are the most significant values found in solenoid datasheets.

Holding forces for DC solenoids can range from a few grams to hundreds of kilograms. The holding force will be inversely proportional to the length of the solenoid, if all other variables are equal. The force that the solenoid can exert on its plunger also varies depending on the position of the plunger in the length of its throw.

Duty cycle is of special importance because the solenoid continues to draw power and create heat so long as it is holding the plunger at the end of its throw (assuming the solenoid is not the latching type). The initial current surge in an AC solenoid generates additional heat.

The duty cycle is simply calculated. If T1 is the time for which the solenoid is on and T2 is the time for which the solenoid is off, the duty cycle, D, is derived as a percentage from the formula

$$D = 100 * (T1 / (T1 + T2))$$

Some solenoids are designed to withstand a 100% duty cycle, but many are not, and in those cases, there is a maximum value not only for D but for the peak “on” time, regardless of the duty cycle. Suppose a solenoid is rated for a 25% duty cycle. If the solenoid is appropriately switched on for one second and off for three seconds, the heat will be allowed to dissipate before it has time to reach overload levels. If the solenoid is switched

on for one minute and off for three minutes, the duty cycle is still 25%, but the heat that may accumulate during a one-minute “on” cycle may overload the component before the “off” cycle can allow it to dissipate.

Coil Size vs. Power

Because additional windings in a coil will induce a greater magnetic force, a larger solenoid tends to be more powerful than a smaller solenoid. However this means that if a larger and a smaller solenoid are both designed to generate the same force over the same distance, the smaller solenoid will probably draw more current (and will therefore generate more heat) because of its fewer coil windings.

How to Use it

Solenoids are primarily used to operate valves in fluid and gas circuits. Such circuits are found in laboratory and industrial process control, fuel injectors, aircraft systems, military applications, medical devices, and space vehicles. Solenoids may also be used in some electronic locks, in pin-ball machines, and in robotics.

What Can Go Wrong

Heat

Overheating is the principal concern when using solenoids, especially if the maximum “on” time is exceeded, or the duty cycle is exceeded. If the plunger is prevented from reaching the end of its throw, this can be another cause of overheating.

Because coil resistance increases with heat, a hot solenoid passes less current and therefore de-

velops less power. This effect is more pronounced in a DC solenoid than an AC solenoid. A manufacturer’s force curve should show the solenoid performance at its maximum rated temperature, which is typically around 75 degrees Centigrade, in a hypothetical ambient temperature of 25 degrees Centigrade. Exceeding these values may result in the solenoid failing to perform. As in all coils using magnet wire, there is the risk of excessive heat melting the insulation separating the coil windings, effectively shortening the coil, which will then pass more current, generating more heat.

AC Inrush

When an AC solenoid reaches the end of its travel, the sudden stop of the plunger results in forward *EMF* that generates additional heat. Generally speaking, a longer stroke creates a greater surge. Rapid cycling will therefore exacerbate coil heating.

Unwanted EMF

Like any device containing a coil, a solenoid creates back EMF when power is connected, and forward EMF when the power is disconnected. A protection diode may be necessary to suppress power spikes that can affect other components.

Loose Plunger

The plunger in many solenoids is not anchored or retained inside the frame and may fall out if the solenoid is tilted or subjected to extreme vibration.

DC motor

In this section, the term “traditional DC motor” is used to describe the oldest and simplest design, which consists of two *brushes* delivering power via a rotating, sectioned *commutator* to two or more electromagnetic coils mounted on the motor shaft. *Brushless* DC motors (in which DC is actually converted to a pulse train) are also described here because “brushless DC” has become a commonly used phrase, and the motor is powered by direct current, even though this is modified internally via *pulse-width modulation*.

OTHER RELATED COMPONENTS

- **AC motor** (See [Chapter 23](#))
- **stepper motor** (See [Chapter 25](#))
- **servo motor** (See [Chapter 24](#))

What It Does

A traditional DC motor uses direct current to create magnetic force, which turns an output shaft. When the polarity of the DC voltage is reversed, the motor reverses its direction of rotation. Usually, the force created by the motor is equal in either direction.

How It Works

Current passes through two or more coils that are mounted on the motor shaft and rotate with it. This assembly is referred to as the *rotor*. The magnetic force produced by the current is concentrated via cores or poles of soft iron or high-silicon steel, and interacts with fields created by permanent magnets arrayed around the rotor in a fixed assembly known as the *stator*.

Power to the coils is delivered through a pair of *brushes*, often made from a graphite compound. Springs press the brushes against a sleeve that rotates with the shaft and is divided into sections, connected with the coils. The sleeve assembly is

known as the *commutator*. As the commutator rotates, its sections apply power from the brushes to the motor coils sequentially, in a simple mechanical switching action.

The most elementary configuration for a traditional DC motor is shown in [Figure 22-1](#).

In reality, small DC motors typically have three or more coils in the rotor, to provide smoother operation. The operation of a three-coil motor is shown in [Figure 22-2](#). The three panels in this figure should be seen as successive snapshots of one motor in which the rotor turns progressively counter-clockwise. The brushes are colored red and blue to indicate positive and negative voltage supply, respectively. The coils are wired in series, with power being applied through the commutator to points between each pair of coils. The direction of current through each coil determines its magnetic polarity, shown as N for north or S for south. When two coils are energized in series without any power applied to their mid-point, each develops a smaller magnetic field than an individually energized coil. This is

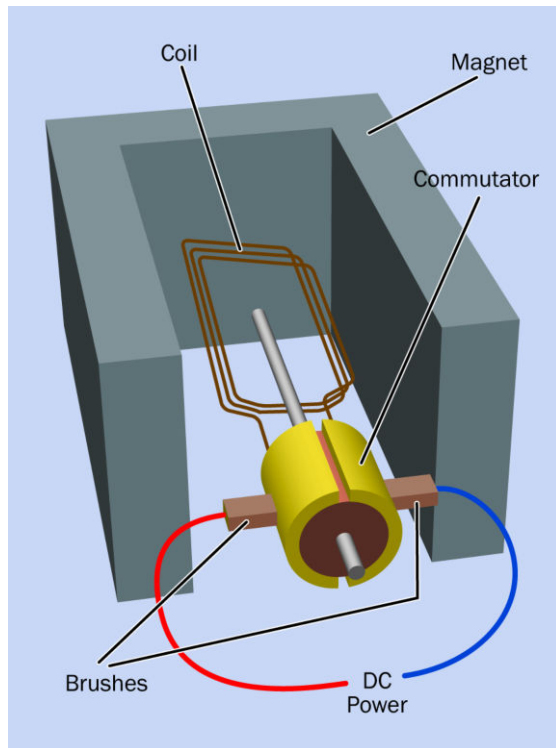


Figure 22-1. The simplest traditional DC motor contains these parts. The combination of coil, shaft, and commutator is the *rotor*. The fixed magnetic structure in which it rotates is the *stator*.

indicated in the diagram with a smaller white lowercase n and s. When two ends of a coil are at equal potential, the coil produces no magnetic field at all.

The stator consists of a cylindrical permanent magnet, which has two poles—shown in the figure as two black semicircles separated by a vertical gap for clarity—although in practice the magnet may be made in one piece. Opposite magnetic poles on the rotor and stator attract each other, whereas the same magnetic poles repel each other.

DC motors may be quite compact, as shown in [Figure 22-3](#), where the frame of the motor measures about 0.7" square. They can also be very powerful for their size; the motor that is shown disassembled in [Figure 22-4](#) is from a 12VDC bilge pump rated at 500 gallons per hour. Its out-

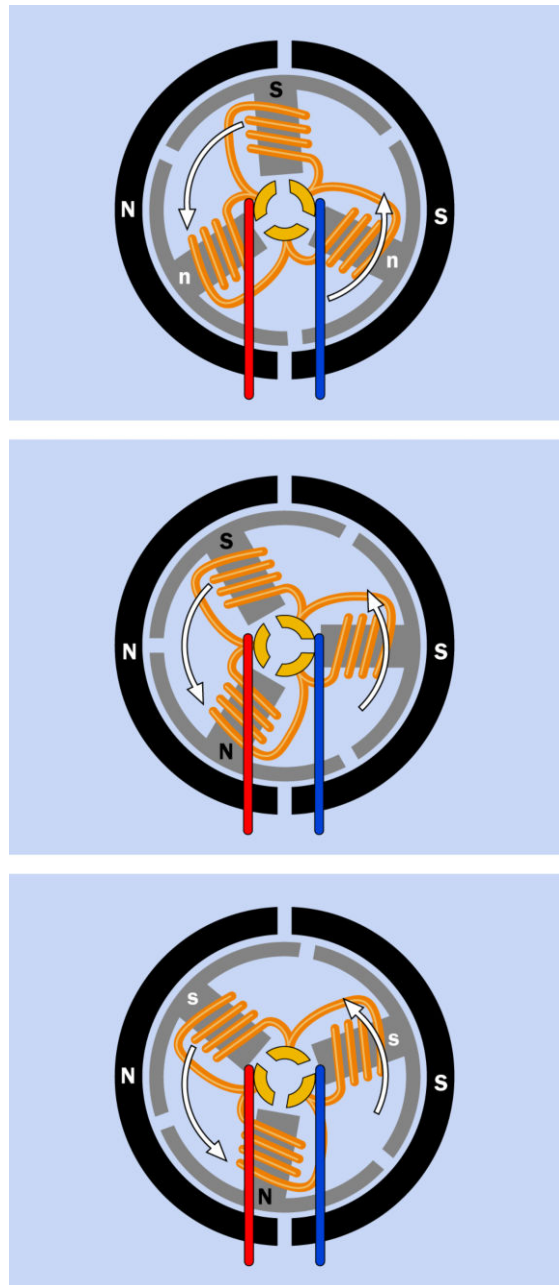


Figure 22-2. Three sequential views of a typical three-coil DC motor viewed from the end of its shaft (the shaft itself is not shown). Magnetic effects cause the rotor to turn, which switches the current to the coils via the commutator at the center.

put was delivered by the small impeller attached to the rotor at right, and was achieved by using

two extremely powerful neodymium magnets, just visible on the inside of the motor's casing (at top-left) in conjunction with five coils on the rotor.

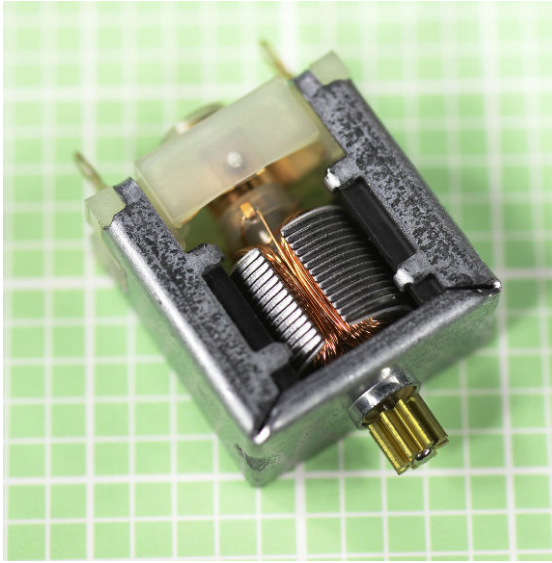


Figure 22-3. A miniature 1.5VDC motor measuring about 0.7" square.

Variants

Coil Configurations

The series connection of coils used in [Figure 22-2](#) is known as the *delta configuration*. The alternative is the *wye configuration* (or *Y configuration*, or *star configuration*). Simplified schematics are shown in [Figure 22-5](#). Generally speaking, the delta configuration is best suited to high-speed applications, but provides relatively low torque at low speed. The wye configuration provides higher torque at low speed, but its top speed is limited.

Gearhead Motor

A *gearhead motor* (also often known as a *gear motor*) incorporates a set of reduction gears that increase the torque available from the output shaft while reducing its speed of rotation. This is often desirable as an efficient speed for a traditional DC motor may range from 3,000 to 8,000

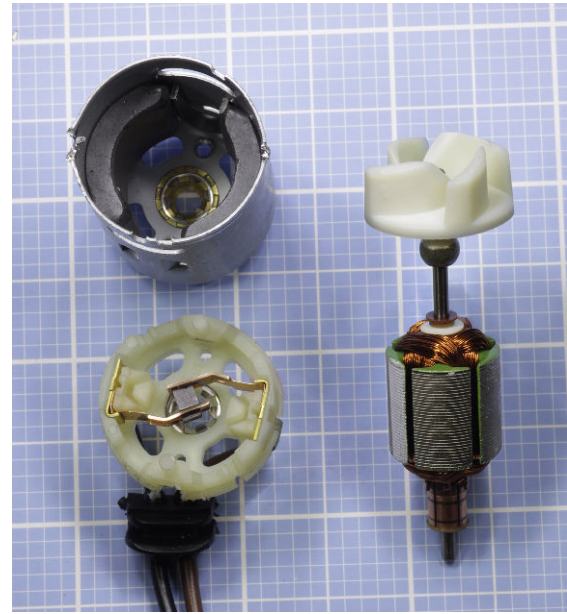


Figure 22-4. A traditional DC motor removed from its cylindrical casing. The brushes of the motor are attached to the white plastic end piece at bottom-left. Large squares on the graph paper in the background are 1" × 1", divided in 0.1" increments. The motor was used in a small bilge pump.

RPM, which is too fast for most applications. The gears and the motor are often contained in a single sealed cylindrical package. Two examples are shown in [Figure 22-6](#). A disassembled motor, revealing half of its gear train under the cap and the other half still attached to a separate circular plate, appears in [Figure 22-7](#). When the motor is assembled, the gears engage. As in the case of the bilge-pump motor, the stator magnets are mounted inside the cylindrical casing. Note that the brushes, inside the circular plate of white plastic, have a resistor and capacitor wired to suppress voltage spikes.

Spur gears are widely used for speed reduction. *Planetary gears* (also known as *epicyclic gears*) are a slightly more expensive option. Spur gears such as those in [Figure 22-8](#) may require three or more pairs in series. The total speed reduction is

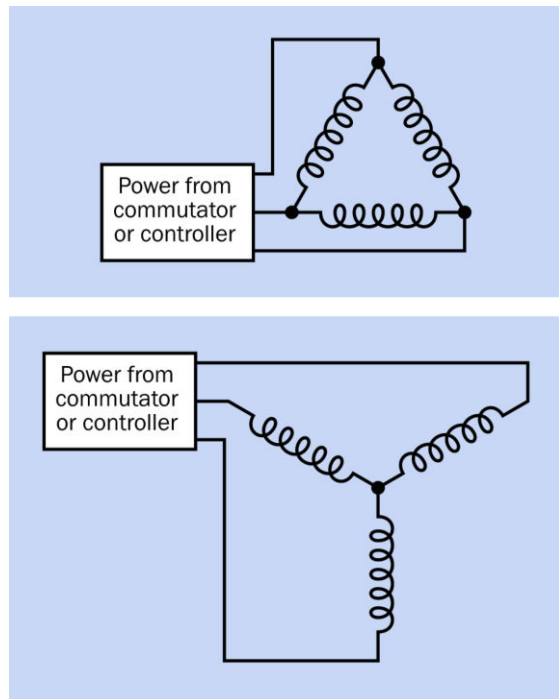


Figure 22-5. Coils on the rotor of a traditional DC motor may be connected in delta configuration (top) or wye configuration (bottom).



Figure 22-6. Two typical small gearhead motors.

found by multiplying the individual ratios. Thus, if three pairs of gears have ratios of 37 : 13, 31 : 15, and 39 : 17, the total speed reduction[®] is obtained by:

$$R = (37 * 31 * 39) / (13 * 15 * 17)$$



Figure 22-7. Spur gears from a gearhead motor provide speed reduction and increased torque.

Therefore:

$$R = 44733 / 3315 = \text{about } 13.5 : 1$$

Datasheets almost always express R as an integer. For example, the gear train shown in [Figure 22-7](#) is rated by the manufacturer as having an overall reduction of 50:1. In reality, the reduction can be expected to have a fractional component. This is because if two gears have an integer ratio, their operating life will be shortened, as a manufacturing defect in a tooth in the smaller gear will hit the same spots in the larger gear each time it rotates. For this reason, the numbers of teeth in two spur gears usually do not have any common factors (as in the example above), and if a motor rotates at 500 RPM, a gear ratio stated as 50:1 is very unlikely to produce an output of exactly 10

RPM. Since traditional DC motors are seldom used for applications requiring high precision, this is not usually a significant issue, but it should be kept in mind.

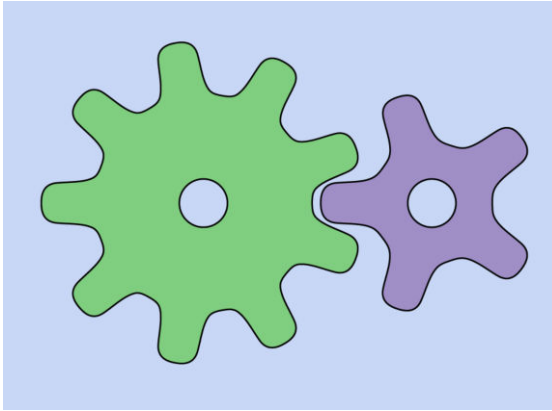


Figure 22-8. A pair of spur gears.

Figure 22-9 shows *planetary gears*, also known as *epicyclic gears*. The outer ring gear is properly referred to as the *annulus*, while the *sun gear* is at the center, and the intermediate *planet gears* may be mounted on a *carrier*. The greatest speed reduction will be achieved by driving the sun gear while the annulus is kept in a stationary position and the output is taken from the carrier of the planet gears. If A is the number of teeth in the annulus and S is the number of teeth in the sun gear, the total speed reduction, R, is given by the following formula:

$$R = (S + A) / S$$

Note that in this drive configuration, the number of teeth in each planet gear is irrelevant to the speed reduction. In Figure 22-9, the sun gear has 27 teeth whereas the annulus has 45 teeth. Therefore, the reduction is found by:

$$R = (27 + 45) / 27 = \text{about } 2.7 : 1$$

Successive reductions can be achieved by stacking planetary gear sets, using the carrier of one set to drive the sun gear in the next set.

Planetary gears are used primarily if a motor drives a heavy load, as the force is divided among

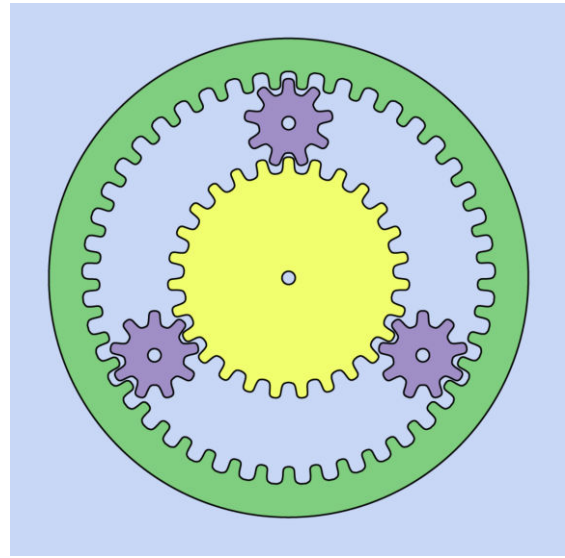


Figure 22-9. Planetary gears, also known as *epicyclic gears*, share the torque from a motor among more teeth than simple spur gears.

more gear pairs, reducing wear and tear on gear teeth and minimizing the breakdown of lubrication. A planetary gear train may also be more compact than a train of spur gears. These advantages must be evaluated against the higher price and slightly increased friction resulting from the larger number of gears interacting with each other.

Brushless DC Motor

In a *brushless* DC motor, sometimes referred to as a *BLDC motor*, the coils are located in the stator and the permanent magnets are relocated in the rotor. The great advantage of this design is that power can be applied directly to the coils, eliminating the need for brushes, which are the primary source of failure in DC motors as a result of wear and tear. However, since there is no rotating commutator to switch the DC current to the coils, the current must be switched by electronic components, which add to the cost of the motor.

In the *inrunner* configuration the stator surrounds the rotor, whereas in the *outrunner* configuration the stator is located in the center of the motor while the rotor takes the form of a ring or

cup that spins around the stator. This is a common design for small cooling fans, where the blades are attached to the outer circumference of a cup that is lined with permanent magnets. An example is shown in [Figure 22-10](#). In this picture, the stator coils are normally hidden from view, being fixed to the fan housing (shown at the top of the picture). Power is controlled by the surface-mount components on the green circular circuit board. The cup attached to the fan blades contains permanent magnets.

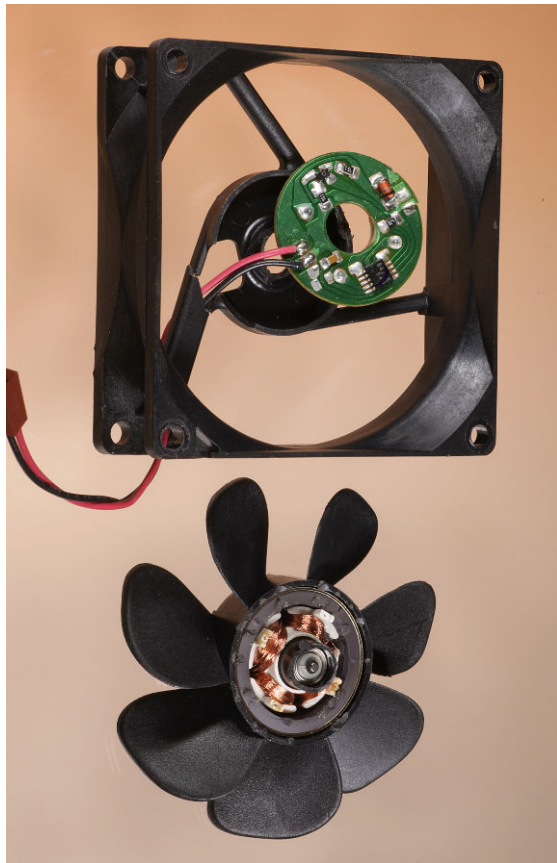


Figure 22-10. A typical brushless DC cooling fan uses stationary coils, with permanent magnets rotating around them.

The use of a solid-state switching system to energize the coils sequentially is known as [electron-ic commutation](#). [Hall effect](#) sensors may be used to detect the position of the rotor and feed this information back to the frequency control circuit,

so that it stays “one step ahead” of the rotor (when bringing it up to speed) or is synchronized with the rotor (for a constant running speed). The system is comparable to a [reluctance motor](#) or [synchronous motor](#). These variants are described in the **AC motor** section of this encyclopedia.

While traditional DC motors have been commercially available since the late 1800s, brushless DC motors were not introduced until the 1960s, when the availability of solid-state control electronics began to make the motor design economically viable.

Linear Actuator

[Linear actuator](#) is a generic term for any device that can exert a pushing or pulling force in a straight line. In industrial applications, actuators may be powered pneumatically or hydraulically, but smaller-scale units are usually driven by a traditional DC motor. These are more properly (but not often) referred to as [electromechanical linear actuators](#).

The rotational force of the motor is typically converted to linear motion by using a threaded motor shaft in conjunction with a nut or collar. The unit is often mounted in an enclosure containing [limit switches](#) that stop the motor automatically at the limits of travel. For an explanation of limit switches, see “[Limit Switches](#)” ([page 46](#)) in the **switch** entry in this encyclopedia.

Values

A manufacturer’s datasheet should list the maximum operating voltage and typical current consumption when a motor is moderately loaded, along with the [stall current](#) that a motor draws when it is so heavily loaded that it stops turning. If stall current is not listed, it can be determined empirically by inserting an ammeter (or multimeter set to measure amperes) in series with the motor and applying a braking force until the motor stops. Motors should generally be protected with slow-blowing **fuses** to allow for the power fluctuations that occur when the motor starts running or experiences a change in load.

In addition, the *torque* that a motor can deliver should be specified. In the United States, torque is often expressed in *pound-feet* (or *ounce-inches* for smaller motors). Torque can be visualized by imagining an arm pivoted at one end with a weight hung on the other end. The torque exerted at the pivot is found by multiplying the weight by the length of the arm.

In the metric system, torque can be expressed as gram-centimeters, Newton-meters, or dyne-meters. A Newton is 100,000 dynes. A dyne is defined as the force required to accelerate a mass of 1 gram, increasing its velocity by 1 centimeter per second each second. 1 Newton-meter is equivalent to approximately 0.738 pound-feet.

The speed of a traditional DC motor can be adjusted by varying the voltage to it. However, if the voltage drops below 50% of the rated value, the motor may simply stop.

The power delivered by a motor is defined as its speed multiplied by its torque at that speed. The greatest power will be delivered when the motor is running at half its *unloaded speed* while delivering half the *stall torque*. However, running a motor under these conditions will usually create unacceptable amounts of heat, and will shorten its life.

Small DC motors should be run at 70% to 90% of their unloaded speed, and at 10% to 30% of the stall torque. This is also the range at which the motor is most efficient.

Ideally, DC motors that are used with reduction gearing should be driven with less than their rated voltage. This will prolong the life of the motor.

When choosing a motor, it is also important to consider the *axial loading* (the weight or force that will be imposed along the axis or shaft of the motor) and *radial loading* (the weight or force that will be imposed perpendicularly to the axis). Maximum values should be found in motor datasheets.

In the hobby field, motors for model aircraft are typically rated in watts-per-pound of motor weight (abbreviated w/lb). Values range from 50 to 250 w/lb, with higher values enabling better performance.

Relationships between torque, speed, voltage, and amperage in a traditional DC motor can be described easily, assuming a hypothetical motor that is 100% efficient:

If the amperage is constant, the torque will also be constant, regardless of the motor speed.

If the load applied to the motor remains constant (thus forcing the motor to apply a constant torque), the speed of the motor will be determined by the voltage applied to it.

If the voltage to the motor remains constant, the torque will be inversely proportional with the speed.

How to Use it

A traditional DC motor has the advantages of cheapness and simplicity, but is only suitable for intermittent use, as its brushes and commutator will tend to limit its lifetime. Its running speed will be approximate, making it unsuitable for precise applications.

As the cost of control electronics has diminished, brushless DC motors have replaced traditional DC motors. Their longevity and controllability provide obvious advantages in applications such as hard disk drives, variable-speed computer fans, CD players, and some workshop tools. Their wide variety of available sizes, and good power-to-weight ratio, have encouraged their adoption in toys and small vehicles, ranging from remote-controlled model cars, airplanes, and helicopters to personal transportation devices such as the Segway. They are also used in direct-drive audio turntables.

Where an application requires the rotation of a motor shaft to be converted to linear motion, a prepackaged *linear actuator* is usually more reliable and simpler than building a crank and

connecting rod, or cam follower, from scratch. Large linear actuators are used in industrial automation, while smaller units are popular with robotics hobbyists and can also be used to control small systems in the home, such as a remote-controlled access door to a home entertainment center.

Speed Control

A *rheostat* or **potentiometer** may be placed in series with a traditional DC motor to adjust its speed, but will be inefficient, as it will achieve a voltage drop by generating heat. Any rheostat must be rated appropriately, and should probably be wire-wound. The voltage drop between the wiper and the input terminal of the rheostat should be measured under a variety of operating conditions, along with the amperage in the circuit, to verify that the wattage rating is appropriate.

Pulse-width modulation (PWM) is preferable as a means of speed control for a traditional DC motor. A circuit that serves this purpose is sometimes referred to as a *chopper*, as it chops a steady flow of current into discrete pulses. Usually the pulses have constant frequency while varying in duration. The pulse width determines the average delivered power, and the frequency is sufficiently high that it does not affect smoothness of operation of the motor.

A **programmable unijunction transistor** or PUT can be used to generate a train of pulses, adjustable with a potentiometer attached to its emitter. Output from the transistor goes to a *silicon-controlled rectifier (SCR)*, which is placed in series with the motor, or can be connected directly to the motor if the motor is small. See [Figure 27-7](#).

Alternatively, a *555 timer* can be used to create the pulse train, controlling a *MOSFET* in series with the motor.

A **microcontroller** can also be used as a pulse source. Many microcontrollers have PWM capability built in. The microcontroller will require its

own regulated power supply (typically 5VDC, 3.3VDC, or sometimes less) and a switching component such as an *insulated-gate bipolar transistor (IGBT)* to deliver sufficient power to the motor and to handle the flyback voltage. These components will all add to the cost of the system, but many modern devices incorporate microcontrollers anyway, merely to process user input. Another advantage of using a microcontroller is that its output can be varied by rewriting software, for example if a motor is replaced with a new version that has different characteristics, or if requirements change for other reasons. Additionally, a microcontroller enables sophisticated features such as pre-programmed speed sequences, stored memory of user preferences, and/or responses to conditions such as excessive current consumption or heat in the motor.

A PWM schematic using a microcontroller and IGBT is shown in [Figure 22-11](#).

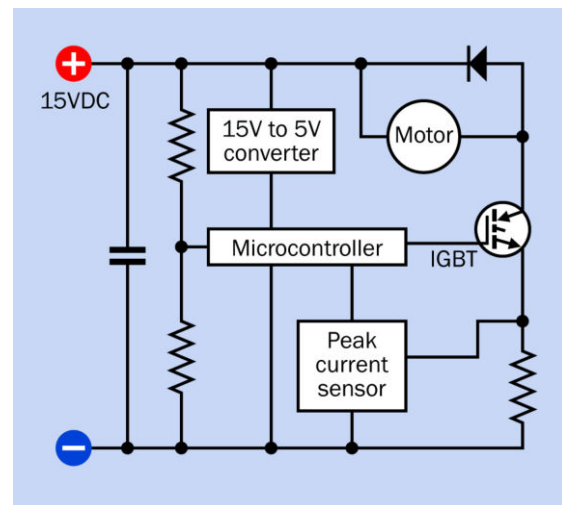


Figure 22-11. A sample schematic for control of a DC motor via pulse-width modulation, using a microcontroller and an insulated-gate bipolar transistor.

Direction Control

The *H bridge* is a very early system for reversing the direction of a DC motor simply by swapping the polarity of its power supply. This is shown in [Figure 22-12](#). The switches diagonally opposite

each other are closed, leaving the other two switches open; and then to reverse the motor, the switch states are reversed. This is obviously a primitive scheme, but the term “H bridge” is still used when prepackaged in a single chip such as the LMD18200 H bridge motor controller from National Semiconductor.

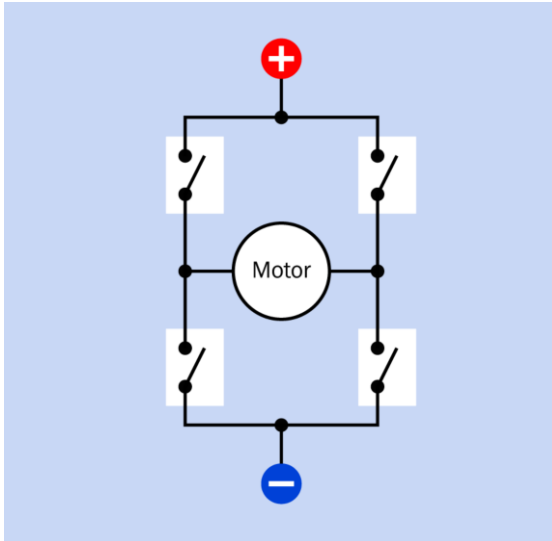


Figure 22-12. A DC motor can be reversed by this very basic circuit, known as an H bridge, by opening and closing pairs of switches that are diagonally opposite each other.

A double-throw, double-pole **switch** or **relay** can achieve the same purpose, as shown in [Figure 22-13](#).

Limit Switches

When a traditional DC motor is used reversibly within a restricted range of motion, it can be fitted with *limit switches* to prevent the motor from stalling and burning out at either end of its permitted travel. Limit switches are explained in “[Limit Switches](#)” ([page 46](#)) in the **switch** entry in this encyclopedia.

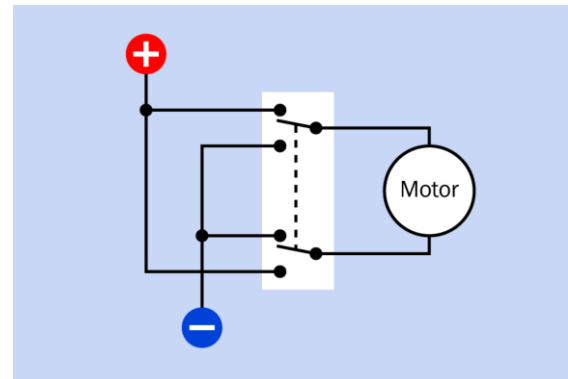


Figure 22-13. A DPDT switch or relay can reverse the direction of a traditional DC motor simply by swapping the polarity of the power supply.

What Can Go Wrong

Brushes and Commutator

The primary cause of failure in DC motors is abrasion of the brushes and wear and tear, oxidation, and/or accumulation of dirt on the commutator. Some motors are designed to allow replacement of the brushes; sealed motors and gearhead motors generally are not. High current and high speed will both tend to accelerate wear in the areas where the brushes meet the commutator.

Electrical Noise

The intermittent contact between the brushes and sections of the commutator of a traditional DC motor can induce voltage spikes that may travel back into the power supply for the motor and cause seemingly random effects in other components. Sparking in the commutator can be a significant source of *electromagnetic interference (EMI)*, especially where cheap or poorly fitted brushes are used. Even if the commutator is running cleanly, the rapid creation of a magnetic field in a motor winding, following by collapse of the field, can create spikes that feed back into the power supply.

Wires that power a motor should be in twisted-pair configuration, so that their radiated EMI tends to cancel itself out. They should be routed

away from data lines or encoder outputs, and may be shielded if necessary. Data lines from sensors in brushless motors may also be shielded.

Installing a capacitor across the motor terminals can significantly reduce EMI. Some motors have capacitors preinstalled. If the motor is in a sealed casing, it may have to be disassembled to reveal whether a capacitor is present.

Heat effects

Since all motors in the real world are less than 100% efficient, some power is lost by the motor during normal operation, and will be dissipated as heat. The resistance of motor windings, and consequently the magnetic force that they generate, will decrease as the temperature rises. The motor becomes less efficient, and will try to draw more current, worsening the situation. A manufacturer's rating for maximum temperature should be taken seriously.

The insulation of the coil windings is usually the most vulnerable part of a motor if excess heat persists. Short circuits between adjacent coils as a result of insulation breakdown will degrade the performance of the motor while increasing its power consumption, which will create even more heat.

Where motor casings have protruding ridges, these cooling fins should be exposed to ambient air.

Frequent starting, stopping, and reversing will tend to generate heat as a result of power surges, and will reduce the lifetime of the motor.

Ambient Conditions

A warm, dry environment will tend to dry out bearing lubricants and graphite brushes. Conversely, a very cold environment will tend to thicken the bearing lubricants. If a motor will be used in unusual environmental conditions, the manufacturer should be consulted.

Wrong Shaft Type or Diameter

Motors have a variety of possible output shaft diameters, some measured in inches and others in millimeters, and shafts may be long or short, or may have a D-shaped cross section or splines to mate with appropriate accessories such as gears, pulleys, or couplings. Careful examination of datasheets and part numbers is necessary to determine compatibility. In the hobby-electronics world, retailers may offer purpose-built discs or arms for specific motor shafts.

Incompatible Motor Mounts

Mounting lugs or flanges may or may not be provided, and may be incompatible with the application for which the motor is intended. The same motor may be available with a variety of mount options, differentiated only by one letter or digit in the motor's part number. A mount option that was available in the past may become obsolete or may simply be out of stock. Again, examination of datasheets is necessary.

Backlash

Backlash is the looseness or "slack" in a gear train that results from small gaps between meshing gear teeth. Because backlash is cumulative when gears are assembled in series, it can become significant in a slow-output gearhead motor. When measured at the output shaft, it is generally in the range of 1 to 7 degrees, tending to increase as the load increases. If a geared motor is used as a positioning device, and is fitted with an encoder to count rotations of the motor shaft, control electronics may cause the motor to hunt to and fro in an attempt to overcome the hysteresis allowed by the backlash. A **stepper motor** or **servo motor** is probably better suited to this kind of application.

Bearings

When using a motor that is not rated for significant axial loading, the bearings may be damaged by applying excessive force to push-fit an output

gear or pulley onto the motor shaft. Even minor damage to bearings can cause significant noise (see the following section) and a reduced lifetime for the component.

In brushless DC motors, the most common cause of failure is the deterioration of bearings. Attempting to revive the bearings by unsealing them and adding lubricant is usually not worth the trouble.

Audible Noise

While electric motors are not generally thought of as being noisy devices, an enclosure can act as a sounding board, and bearing noise is likely to increase over time. Ball bearings become noisy over time, and gears are inherently noisy.

If a device will contain multiple motors, or will be used in close proximity to people who are likely to be sensitive to noise (for example, in a medical environment), care should be taken to insure that motor shafts are properly balanced, while the motors may be mounted on rubber bushings or in sleeves that will absorb vibration.

AC motor

The distinction between AC and DC motors has become blurred as controllers for DC motors make increasing use of *pulse-width modulation*, which can be viewed as a form of alternating current. All motors that consume DC power are referenced in the **DC motor** section of this encyclopedia, regardless of whether they modulate the power internally. **Stepper motors** and **servo motors** are considered as special cases, each with its own entry. AC motors, described here, are those that consume alternating current, usually in the form of a sine wave with a fixed frequency.

OTHER RELATED COMPONENTS

- **DC motor** (see [Chapter 22](#))
- **stepper motor** (see [Chapter 25](#))
- **servo motor** (see [Chapter 24](#))

What It Does

An AC motor uses a power supply of alternating current to generate a fluctuating magnetic field that turns a shaft.

How It Works

The motor consists primarily of two parts: the *stator*, which remains stationary, and the *rotor*, which rotates inside the stator. Alternating current energizes one or more coils in the stator, creating fluctuating magnetic fields that interact with the rotor. A simplified representation is shown in [Figure 23-1](#), where the coils create magnetic forces indicated by the green arrows, N representing North and S representing South.

Stator Design

Plug-in electric fans typically use AC motors. The stator from a large electric fan is shown in [Figure 23-2](#), where the large diameter of each coil

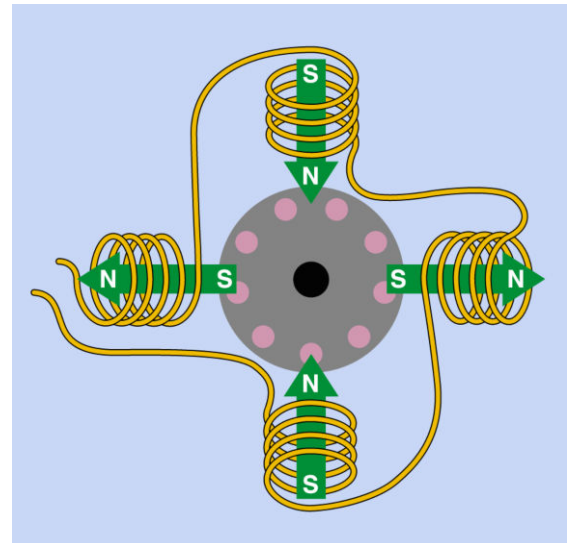


Figure 23-1. A simplified representation of a basic AC motor. The green arrows indicate magnetic force.

maximizes its magnetic effect. The stator from a smaller electric fan is shown in [Figure 23-3](#), in which only one coil is used. (The coil is wrapped in black electrical tape.)



Figure 23-2. The stator from a large electric fan. Each coil of copper wire is centered on a lug pointing inward to the hole at the center, where the rotor would normally be mounted. The coils overlap because their diameter is maximized to increase their magnetic effect. Each coil is tapped to allow speed selection in steps via an external rotary switch.

The core of a stator resembles the core of a **transformer** in that it usually consists of a stack of wafers of high-silicon steel (or sometimes aluminum or cast iron). The layers are insulated from one another by thin layers of shellac (or a similar compound) to prevent eddy currents that would otherwise circulate through the entire thickness of the stator, reducing its efficiency.

The coil(s) wound around the stator are often referred to as *field windings*, as they create the magnetic field that runs the motor.

Rotor Design

In most AC motors, the rotor does not contain any coils and does not make any electrical con-

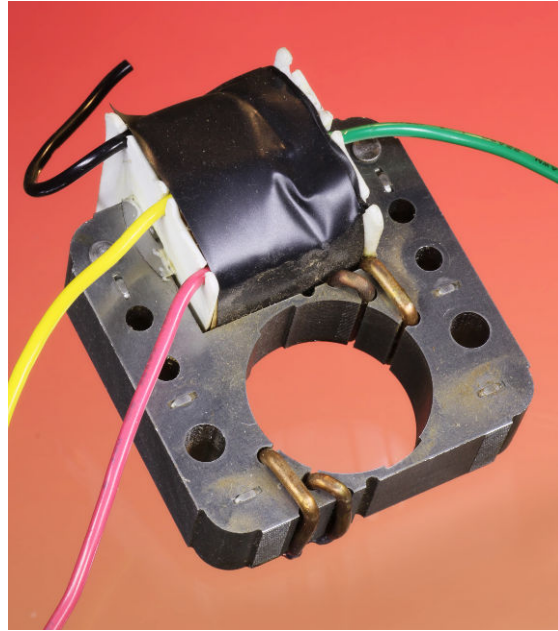


Figure 23-3. This stator from a smaller electric fan uses only a single coil, wrapped in black tape. It is sufficient to induce a magnetic field but is generally less efficient than a motor using multiple coils.

nection with the rest of the motor. It is powered entirely by induced magnetic effects, causing this type of motor to be known generally as an *induction motor*.

As the AC voltage changes from positive to negative, the magnetic force induced in the stator collapses and a new field of opposite polarity is created. Because the stator is designed to create an asymmetrical field, it induces a rotating magnetic field in the rotor. The concept of a rotating magnetic field is fundamental in AC motors.

Like the stator, the rotor is fabricated from wafers of high-silicon steel; embedded in the wafers are nonmagnetic rods, usually fabricated from aluminum but sometimes from copper, oriented approximately parallel to the axis of rotation. The rods are shorted together by a ring at each end of the rotor, forming a conductive “cage,” which explains why this device is often referred to colloquially as a *squirrel cage* motor.

Figure 23-4 shows the configuration of a rotor cage with the surrounding steel wafers removed for clarity. In reality, the rods in the cage are almost always angled slightly, as shown in Figure 23-5, to promote smooth running and reduce *cogging*, or fluctuations in torque, which would otherwise occur.

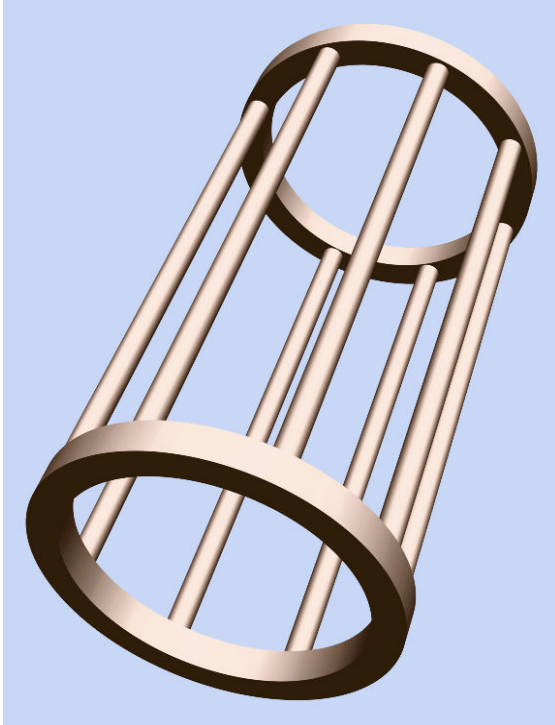


Figure 23-4. The rotor of a typical AC motor contains a cage of aluminum (or sometimes copper) in which eddy currents occur, as a result of the rotating magnetic field inside the steel body of the rotor (which is omitted here for clarity). These currents cause their own magnetic fields, which interact with the fields generated by coils in the stator.

In Figure 23-6, the steel wafers of a rotor are shown, with channels to accommodate an angled aluminum cage. Figure 23-7 shows a cross-section of a rotor with the cage elements in pale red and the steel wafer in gray.

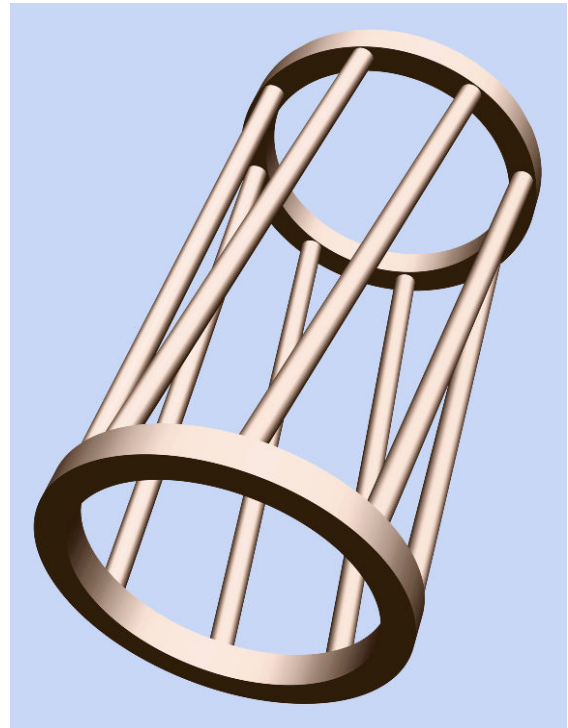


Figure 23-5. To promote smooth running of the motor, the longitudinal elements of the cage are typically angled, as suggested in this rendering.

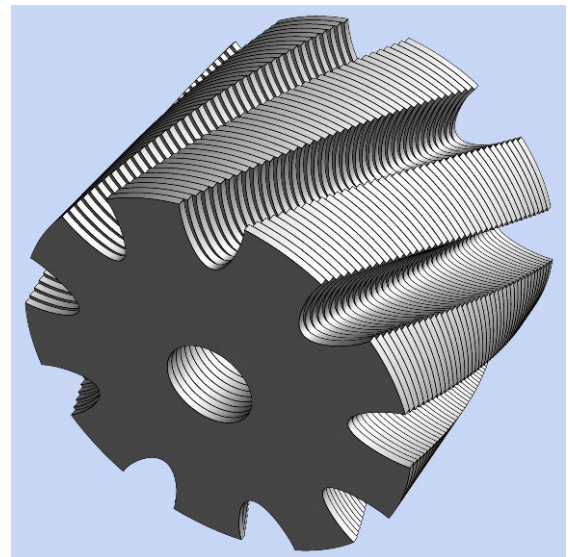


Figure 23-6. The steel wafers in the rotor of an AC motor are typically offset as shown here. The channels are to accommodate a cage of aluminum or copper conductors.

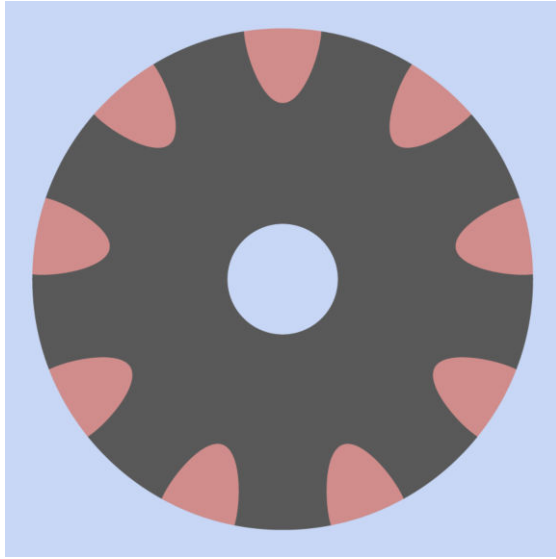


Figure 23-7. Cross-section of a rotor with steel shown in gray and embedded elements of an aluminum cage shown in pale red.

The actual rotor from an induction motor is shown in [Figure 23-8](#). This rotor was removed from the stator shown in [Figure 23-3](#). The bearings at either end of the rotor were bolted to the stator until disassembly.

Although the cage is nonmagnetic, it is electrically conductive. Therefore the rotating magnetic field that is induced in the steel part of the rotor generates substantial secondary electric current in the cage, so long as the magnetic field inside the rotor is turning faster than the rotor itself. The current in the longitudinal elements of the cage creates its own magnetic field, which interacts with the fields created by coils in the stator. Attraction and repulsion between these fields causes the rotor to turn.

Note that if the turning speed of the rotor rises to match the frequency of the alternating current powering the coils in the stator, the cage in the rotor is no longer turning through magnetic lines of force, and ceases to derive any power from



Figure 23-8. The rotor from a small fan motor. The aluminum cage and its end pieces are the pale gray sections, steel plates are the darker sections.

them. In an ideal, frictionless motor, its unloaded operating speed would be in equilibrium with the AC frequency. In reality, an induction motor never quite attains that speed.

When power is applied while the rotor is at rest, the induction motor draws a heavy surge of current, much like a short-circuited **transformer**. Electrically, the coils in the stator are comparable to the primary winding of a transformer, while the cage in the rotor resembles the secondary winding. The turning force induced in the stationary rotor is known as *locked-rotor torque*. As the motor picks up speed, its power consumption diminishes. See [Figure 23-9](#).

When the motor is running and a mechanical load is applied to it, the motor speed will drop. As the speed diminishes, the cage of conductors embedded in the rotor will derive more power, as they are turning more slowly than the rotating magnetic field. The speed of rotation of the field is determined by the frequency of the AC power, and is therefore constant. The difference in rota-

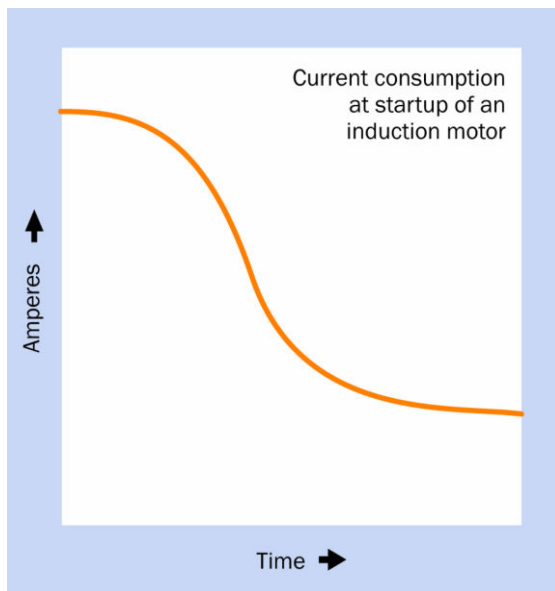


Figure 23-9. An approximated graph showing the typical current consumption of an AC induction motor as it starts from rest and gains speed over a period of time.

tional speed between the magnetic field and the rotor is known as *slip*. Higher levels of slip induce greater power, and therefore the induction motor will automatically find an equilibrium for any load within its designed range.

When running under full load, a small induction motor may have a slip value from 4 to 6 percent. In larger motors, this value will be lower.

Variants

Variants of the generic induction motor described above are generally designed to take advantage of either *single-phase* or *three-phase* alternating current.

A *synchronous motor* is a variant in which the rotor maintains a constant speed of rotation regardless of small fluctuations in load.

Some AC motors incorporate a *commutator*, which allows an external connection to coils mounted on the rotor, and can enable variable speed control.

A *linear motor* may consist of two rows of coils, energized by a sequence of pulses that can move a permanent magnet or electromagnet between the coils. Alternatively, the linear motor's coils may move as a result of magnetic interaction with a segmented fixed rail. Detailed description of linear motors is outside the scope of this encyclopedia.

Single-Phase Induction Motor

The majority of induction motors run on single-phase alternating current (typically, from domestic wall outlets). This type of motor is not innately self-starting because the stator coils and rotor are symmetrical. This tends to result in vibration rather than rotation.

To initiate rotation, the stator design is modified so that it induces an asymmetrical magnetic field, which is more powerful in one direction than the other. The simplest way to achieve this is by adding one or more *shorting coils* to the stator. Each shorting coil is often just a circle of heavy-gauge copper wire. This ploy reduces the efficiency of the motor and impairs its starting torque, and is generally used in small devices such as electric fans, where low-end torque is unimportant. Because the shorting coil obstructs some of the magnetic field, this configuration is often known as a *shaded pole* motor.

Copper shorting coils are visible in the fan motor shown in [Figure 23-3](#).

A **capacitor** is a higher-cost but more efficient alternative to a shorting coil. If power is supplied to one or more of the stator coils through a capacitor, it will create a phase difference between these coils and the others in the motor, inducing an asymmetrical magnetic field. When the motor reaches approximately 80% of its designed running speed, a centrifugal switch may be included to take the capacitor out of the circuit, since it is no longer necessary. Switching out the capacitor and substituting a direct connection to the stator coils will improve the efficiency of the motor.

A third option to initiate rotation is to add a second winding in the stator, using fewer turns of smaller-gauge wire, which have a higher resistance than the main winding. Consequently the magnetic field will be angled to encourage the motor to start turning. This configuration is known as a *split-phase* induction motor, in which the starter winding is often referred to as the *auxiliary winding* and consists of about 30% of the total stator windings in the motor. Here again, a centrifugal switch can be incorporated, to eliminate the secondary winding from the circuit when the motor has reached 75 to 80 percent of its designed running speed.

The relationship between motor speed and torque of the three types of motors described above is shown in Figure 23-10. These curves are simplified and do not show the effect that would be produced by introducing a centrifugal switch.

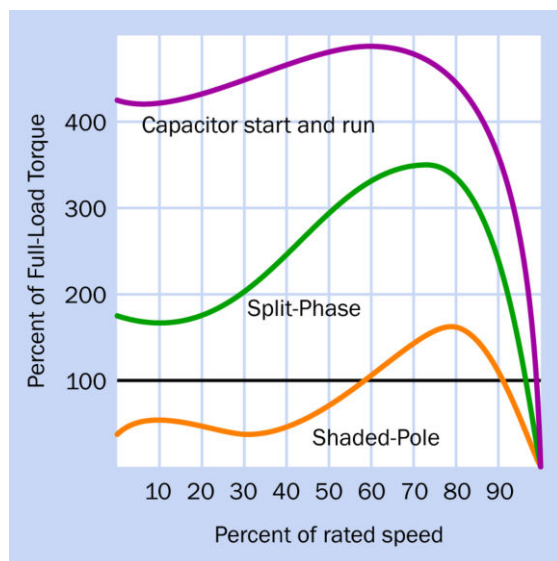


Figure 23-10. Approximate curves showing the relationship between speed and torque for three types of single-phase induction motor. (Graph derived from *AC Induction Motor Fundamentals* published by Microchip Technology Inc.)

Three-Phase Induction Motor

Larger induction motors are often *three-phase* devices. Three-phase AC (which is by far the most

common form of *polyphase* AC) is delivered by a power utility company or generator via three wires, each of which carries alternating current with a phase difference of 120 degrees relative to the other two, usually for industrial applications. A common configuration of stator coils for a three-phase motor is shown in Figure 23-11. Since the three wires take it in turns to deliver their peak voltage, they are ideally suited to turn the stator of a motor via induction, and no shorting coil or capacitor is needed for startup. Heavy-duty 3-phase induction motors are extremely reliable, being brushless and generally maintenance-free.

Synchronous Motor

A synchronous motor is a form of induction motor that is designed to reach and maintain equilibrium when the rotor is turning in perfect synchronization with the AC power supply. The speed of the motor will depend on the number of poles (magnetic coils) in the stator, and the number of phases in the power supply. If R is the RPM of a synchronous motor, f is the frequency of the AC current in Hz, and p is the number of poles per phase:

$$R = (120 * f) / p$$

This formula assumes 60Hz AC current. In nations where 50Hz AC is used, the number 120 should be replaced with the number 100.

Two basic types of synchronous motors exist: *direct current excited*, which require external power to start turning, and *non-excited*, which are self-starting. Since non-excited synchronous motors are more common in electronic applications, this encyclopedia will not deal with direct current excited variants.

A *hysteresis motor* is a synchronous motor containing a solid rotor cast from cobalt steel, which has high coercivity, meaning that once it is magnetized, a substantial field is required to reverse the magnetic polarity. Consequently the polarity of the rotor lags behind the constantly changing

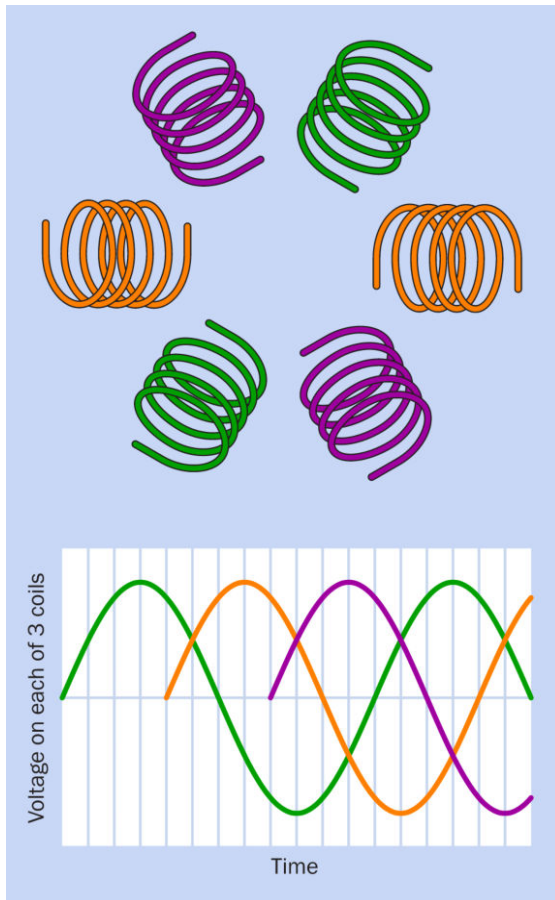


Figure 23-11. The graph shows voltage delivered via three wires constituting a three-phase power supply. (The curve colors are arbitrary.) A three-phase motor contains a multiple of three coils—often six, as shown here diagrammatically. The three wires of the power supply are connected directly to the coils, which induce a rotating magnetic field.

polarity of the stator, creating an attracting force that turns the rotor. Because the lag angle is independent of motor speed, this motor delivers constant torque from startup.

Reluctance Motor

Reluctance is the magnetic equivalent to electrical resistance. If a piece of iron is free to move in a magnetic field, it will tend to align itself with the field to reduce the reluctance of the magnetic

circuit. This principle was used in very early reluctance motors designed to work from AC and has been revived as electronics to control variable frequency drives have become cheaper.

The simplest reluctance motor consists of a soft iron rotor with projecting lugs, rotating within a stator that is magnetically energized with its own set of inwardly projecting poles. The rotor tends to turn until its lugs are aligned with the poles of the stator, thus minimizing the reluctance.

A basic reluctance motor design is shown in [Figure 25-2](#). It is located in the **stepper motor** section of this encyclopedia, as stepper motors are a primary application of the reluctance principle.

Although a reluctance motor can be used with polyphase fixed-frequency AC power, a variable frequency drive greatly enhances its usefulness. The timing of the frequency is adjusted by the speed of the motor, which is detected by a sensor. Thus the energizing pulses can remain “one step ahead” of the rotor. Since the rotor is not a magnet, it generates no back-EMF, allowing it to reach very high speeds.

The simplicity of the motor itself is a compensating factor for the cost of the electronics, as it requires no commutator, brushes, permanent magnets, or rotor windings. Characteristics of reluctance motors include:

- Cheap parts, easily manufactured, and high reliability.
- Compact size and low weight.
- Efficiencies greater than 90% possible.
- Capable of high start-up torque and high speed operation.

Disadvantages include noise, cogging, and tight manufacturing tolerances, as the air gap between the rotor and stator must be minimized.

A reluctance motor can function synchronously, if it is designed for that purpose.

Variable Frequency Drive

A basic induction motor suffers from significant problems. The surge of power that it draws when starting from rest can pull down the supply voltage enough to affect other devices that share the AC power supply. (Hence, the brief dimming of lights that may occur when the compressor in an air conditioner or refrigerator starts running.) While the motor is turning, it can introduce electrical noise, which feeds back into the power supply, once again causing potential problems for other devices. In addition, the narrow range of speed of an AC induction motor is a great disadvantage.

The advent of cheap solid-state technology encouraged the development of *variable-frequency power supplies* for induction motors. Because the impedance of the motor will diminish as the frequency diminishes, the current drawn by the motor will tend to increase. To prevent this, a variable frequency supply also varies the voltage that it delivers.

Wound-Rotor AC Induction Motor

The stator of this variant is basically the same as that of a single-phase induction motor, but the rotor contains its own set of coils. These are electrically accessible via a *commutator* and *brushes*, as in a **DC motor**. Because the maximum torque (also known as *pull-out* torque) will be proportional to the electrical resistance of the coils in the rotor, the characteristics of the motor can be adjusted by adding or removing resistance externally, via the commutator. A higher resistance will enable greater torque at low speed when the slip between the rotor speed and rotation of the magnetic field induced by the stator is greatest. This is especially useful in corded power tools such as electric drills, where high torque at low speed is desirable, yet the motor can accelerate to full speed quickly when the external resistance is reduced. Typically, the resistance is adjusted via the trigger of the drill.

Figure 23-12 shows a wound-rotor AC induction motor. The disadvantage of this configuration is

the brushes that supply power to the rotor will eventually require maintenance. Much larger wound-rotor motors are also used in industrial applications such as printing presses and elevators, where the need for variable speed makes a simple three-phase motor unsuitable.



Figure 23-12. A motor in a corded electric drill uses coils in a brushed rotor to enable variable speed output. In most AC motors, the speed is not adjustable and the rotor does not make any electrical connection with the rest of the motor.

Universal Motor

A wound-rotor motor may also be described as a *universal motor* if its rotor and stator coils are connected in series. This configuration allows it to be powered by either AC or DC.

DC supplied to the rotor and the stator will cause mutual magnetic repulsion. When the rotor

turns, the brushes touching the split commutator reverse the polarity of voltage in the rotor coils, and the process repeats. This configuration is very similar to that of a conventional DC motor, except that the stator in a universal motor uses electromagnets instead of the permanent magnets that are characteristic of a **DC motor**.

When powered by AC, the series connection between stator and rotor coils insures that each pulse to the stator will be duplicated in the rotor, causing mutual repulsion. The addition of a shorting coil in the stator provides the necessary asymmetry in the magnetic field to make the motor start turning.

Universal motors are not limited by AC frequency, and are capable of extremely high-speed operation. They have high starting torque, are compact, and are cheap to manufacture. Applications include food blenders, vacuum cleaners, and hair dryers. In a workshop, they are found in routers and miniature power tools such as the Dremel series.

Because a universal motor requires commutator and brushes, it is only suitable for intermittent use.

Inverted AC Motors

Some modern domestic appliances may seem to contain an AC motor, but in fact the AC current is *rectified* to DC and is then processed with pulse-width modulation to allow variable speed control. The motor is really a **DC motor**; see the entry on this type of motor for additional information.

Values

Because a basic AC induction motor is governed by the frequency of the power supply, the speed of a typical four-pole motor is limited to less than 1,800 RPM (1,500 RPM in nations where 50Hz AC is the norm).

Variable-frequency, universal, and wound-rotor motors overcome this limitation, and can reach

speeds of 10,000 to 30,000 RPM. Synchronous motors typically run at 1,800 or 1,200 RPM, depending on the number of poles in the motor. (They run at 1,500 or 1,000 RPM in locations where the frequency of AC is 50Hz rather than 60Hz).

For a discussion of the torque that can be created by a motor, see “[Values](#)” (page 184) in the **DC motor** entry in this encyclopedia.

How to Use It

Old-fashioned record players (where a turntable supports a vinyl disc that must rotate at a fixed speed) and electric clocks (of the analogue type) were major applications for synchronous motors, which used the frequency of the AC power supply to control motor speed. These applications have been superseded by CD players (usually powered by brushless DC motors) and digital clocks (which use crystal oscillators).

Many home appliances continue to use AC-powered induction motors. Small cooling fans for use in electronic equipment are sometimes AC-powered, reducing the current that must be provided by the DC power supply. An induction motor generally tends to be heavier and less efficient than other types, and its speed limit imposed by the frequency of the AC power supply is a significant disadvantage.

A simple induction motor cannot provide the sophisticated control that is necessary in modern devices such as CD or DVD players, ink-jet printers, and scanners. A **stepper motor**, **servo motor**, and **DC motors** controlled with *pulse-width modulation* are preferable in these applications.

A *reluctance motor* may find applications in high-speed, high-end equipment including vacuum cleaners, fans, and pumps. Large variable reluctance motors, with high amperage ratings, may be used to power vehicles. Smaller variants are being used for power steering systems and windshield wipers in some automobiles.

What Can Go Wrong

Compared with other devices that have moving parts, the brushless induction motor is one of the most reliable and efficient devices ever invented. However, there are many ways it can be damaged. General problems affecting all types of motors are listed at [“Heat effects” \(page 188\)](#). Issues relating specifically to AC motors are listed below.

Premature Restart

Large industrial three-phase induction motors can be damaged if power is reapplied before the motor has stopped rotating.

Frequent Restart

If a motor is stopped and started repeatedly, the heat that is generated during the initial surge of current is likely to be cumulative.

Undervoltage or Voltage Imbalance

A voltage drop can cause the motor to draw more current than it is rated to handle. If this situation persists, overheating will result. Problems also are caused in a three-phase motor where one phase is not voltage-balanced with the others. The most common cause of this problem is an open circuit-breaker, wiring fault, or blown fuse affecting just one of the three conductors. The motor will try to run using the two conductors that are still providing power, but the result is likely to be destructive.

Stalled Motor

When power is applied to an induction motor, if the motor is prevented from turning, the conductors in the rotor will carry a large current that is entirely dissipated as heat. This current surge will either burn out the motor or blow a fuse or circuit breaker. Care should be taken, in equipment design, to minimize the risk that an induction motor may stall or jam.

Protective Relays

Sophisticated protective relays are available for industrial 3-phase motors, and can guard against all of the faults itemized above. Their details are outside the scope of this encyclopedia.

Excess Torque

As has been previously noted, the torque of an induction motor increases with the slip (speed difference) between the rotation of the magnetic field and the rotation of the rotor. Consequently, if the motor is overloaded and forced to run more slowly, it can deliver more rotational force. This can destroy other parts attached to the motor, such as drive belts.

Internal Breakage

An overloaded induction motor may suffer some cracking or breakage of its rotor. This may be obvious because of reduced power output or vibration, but can also be detected if the motor's power consumption changes significantly.

servo motor

Should be referred to as an *RC servo* if it is intended for use in small devices that are remote-controlled and battery powered. However, in practice, the RC acronym is often omitted.

OTHER RELATED COMPONENTS

- **AC motor** (See [Chapter 23](#))
- **DC motor** (See [Chapter 22](#))
- **stepper motor** (See [Chapter 25](#))

What It Does

A servo motor is actually a combination of a motor, reduction gearing, and miniaturized control electronics, usually packaged together inside a very compact sealed plastic case. The motor itself may be AC or DC, and if DC, it may be brushed or brushless. What distinguishes a servo from other types of motor is that it is not designed for continuous rotation. It is a position-seeking device. Its rotational range may be more than 180 degrees but will be significantly less than 360 degrees. Two typical RC servos are shown in [Figure 24-1](#). A side view of a motor is shown in [Figure 24-2](#).

The electronics inside the motor enclosure interpret commands from an external controller. The command code specifies the desired turn angle measured as an offset either side of the center position of the motor's range. The motor turns quickly to the specified position and stops there. So long as the command signal continues and power to the motor is sustained, the motor will hold its position and "push back" against any external turning force. In the absence of such a force, while the motor is stationary, it will use very little current.



Figure 24-1. A typical RC servo motor is capable of more than 50 inch-ounces of torque yet can be driven by three or four AA alkaline cells in series, and weighs under 2 ounces.

The electronics inside a typical RC servo motor are shown in [Figure 24-3](#).

How It Works

Servo motors are generally controlled via *pulse-width modulation (PWM)*.

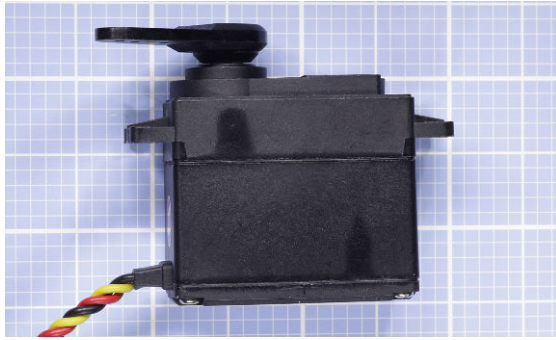


Figure 24-2. RC servo motors are mostly similar in size. This is a typical side view.

An industrial servo typically requires a controller that is an off-the-shelf item sold by the manufacturer of the motor. The encoding scheme of the control signals may be proprietary. A heavy-duty servo may be designed to run from three-phase power at a relatively high voltage, and may be used in applications such as production-line automation.

The remainder of this encyclopedia entry will focus primarily on small RC servos rather than industrial servos.

For small RC servos, the stream of control pulses is at a constant frequency of 20ms, with the positive durations of each pulse being interpreted as a positioning command to the motor, and the gaps between the pulses being disregarded. A typical range of pulse widths for a small motor is 1ms to 2ms, specifying a range of -90 to +90 degrees either side of a center location. Many modern motors are capable of excursions beyond these limits, and can be calibrated to establish the precise relationship between pulse width and turn angle. The motor can then be controlled by a lookup table in microcontroller software, or by using a conversion factor between degree-angle and pulse width.

Figure 24-4 shows the typical range of pulse widths within the fixed 20ms period (a frequency of 50Hz) between the start of one pulse and the

start of the next, and the meaning of each pulse width to the servo motor. Intermediate pulse widths are interpreted as instructions to rotate to intermediate positions.

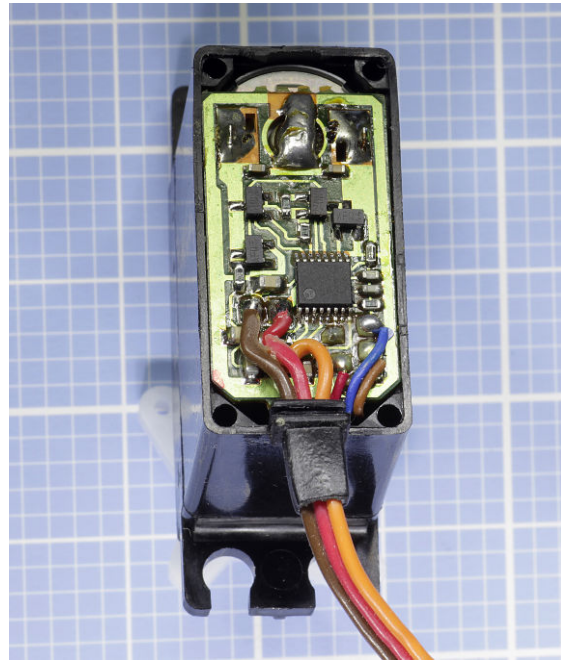


Figure 24-3. The electronics inside a servo motor decode a stream of pulses that specify the turn angle of the motor.

Small servo motors require the user to provide a controller that will conform with the above specification. This is often achieved by programming a [microcontroller](#), and some microcontroller chips make this especially easy by providing a PWM output specifically tailored to the requirements of an RC servo. Either way, the microcontroller can be directly connected to the servo, enabling an extremely simple and flexible way to manage a positioning device.

Alternatively, a simple pulse generator such as a [555 timer chip](#) can be used, or controller boards are available from hobbyist supply sources. Some controller boards have USB connections enabling a servo to be governed by computer software.

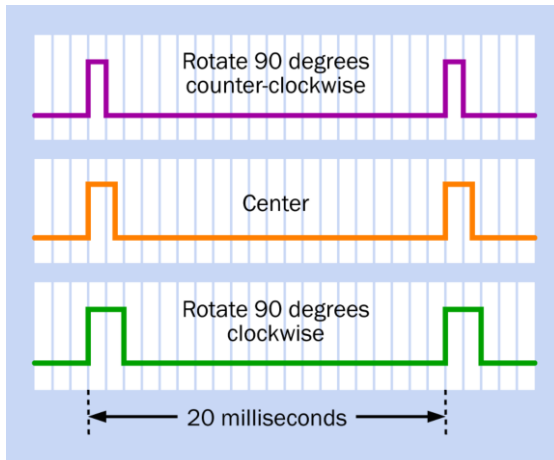


Figure 24-4. The turn angle of a small RC servo motor is determined by a pulse width from a controller ranging from 1ms to 2ms in duration. The frequency of the pulses is constant at 50Hz.

In [Figure 24-5](#), a schematic illustrates the connection of a 555 timer with an RC servo, with component values to create a constant frequency of about 48Hz (slightly more than 20ms from peak to peak). The 1μF capacitor in the circuit charges through the 2.2K resistor in series with the diode, which bypasses the 28K resistor. This charging time represents the “on” cycle of the timer chip. The capacitor discharges through the 28K resistor, representing the “off” cycle. The 1K potentiometer, in series with the 5K resistors, acts as a voltage divider applied to the control pin of the timer, adjusting the timer’s charge and discharge thresholds. Turning the potentiometer will lengthen or shorten the “on” time of each cycle, without changing the frequency. In practice, because capacitors are manufactured with wide tolerances, the frequency of the timer output cannot be guaranteed. Fortunately most servos will tolerate some inaccuracy.

Since the motor shares the power supply of the timer in this circuit, a protection diode and capacitor have been added between the power supply to the motor and negative ground, to suppress noise and back-EMF.

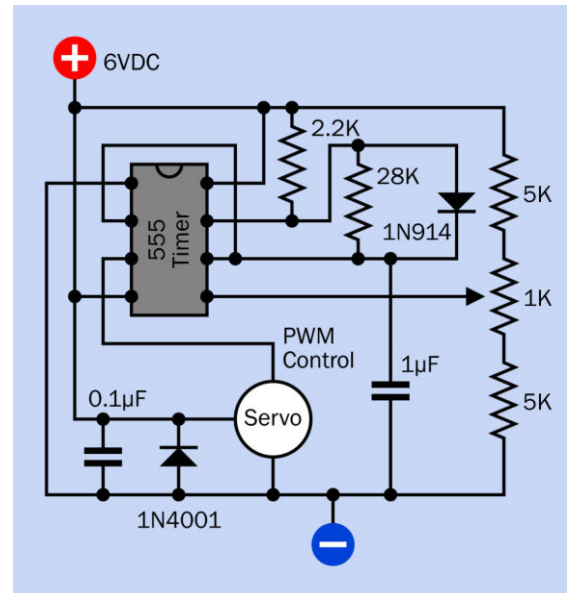


Figure 24-5. An RC servo can be controlled via a 555 timer with appropriate component values. The potentiometer determines the angular position of the servo.

Inside a servo motor’s casing, the electronics include a **potentiometer** that turns with the output shaft, to provide feedback confirming the motor’s position. The limited turning range of the potentiometer determines the turn limits of the motor output shaft.

Variants

Small servos may contain brushed or brushless DC motors. Naturally the brushless motors have greater longevity and create less electrical noise. See the **DC motor** entry in this encyclopedia for a discussion of brushed versus brushless motors.

Servos may use nylon, “Karbonite,” or metal reduction gearing. The nylon gears inside a cheaper RC servo are shown in [Figure 24-6](#).

Brushless motors and metal gears add slightly to the price of the motor. Metal gears are stronger than nylon (which can crack under load) but may wear faster, leading to *backlash* and inaccuracy in the gear train. The friction between nylon-to-nylon surfaces is very low, and nylon is certainly

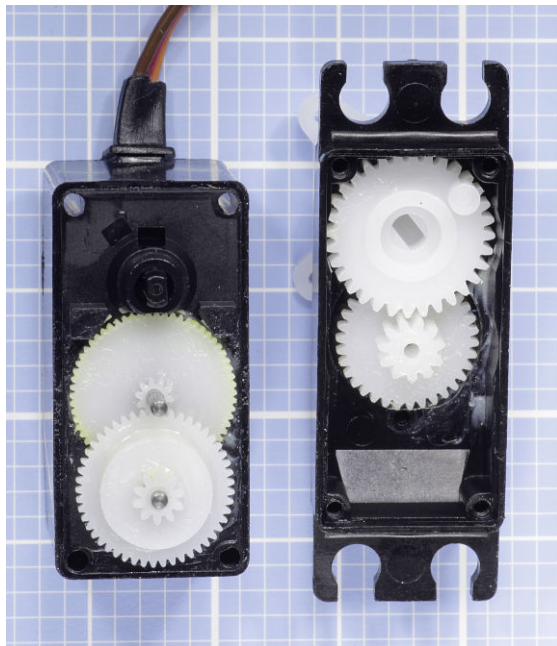


Figure 24-6. Nylon gearing inside a servo motor.

adequate and may be preferable if a servo will not be heavily loaded. “Karbonite” is claimed to be five times stronger than nylon and may be a satisfactory compromise. If a gear set experiences a failure (for example, teeth can be stripped as a result of excessive load), manufacturers usually will sell a replacement set to be installed by the user. Installation requires manual dexterity and patience, and some skill.

Servos may have roller bearings or plain sintered bearings, the latter being cheaper but much less durable under side loading.

So-called *digital servos* use faster internal electronics than the older, so-called *analog servos*, and because they sample the incoming pulse stream at a higher frequency, they are more responsive to small, rapid commands from the controller. For this reason they are preferred by hobbyists using servos to control the flight of model airplanes. Externally, the control protocol for digital and analog servos is the same, although a

digital servo can be reprogrammed with new code values establishing the limits to its range. A standalone programming unit must be purchased to achieve this.

The most popular manufacturers of small servo motors are Futaba and Hitec. While their control protocols are virtually identical, the motor output shafts differ. The shaft is typically known as the *spline*, and is grooved to fit push-on attachments. The spline of a Futaba motor has 25 grooves, while Hitec uses 24 grooves. Attachments must be appropriate for the brand of motor that has been chosen.

Values

A small servo typically weighs 1 to 2 ounces, has a rotation time of 1 to 2 seconds from one end of its travel to the other, and can exert a surprisingly robust torque of 50 ounce-inches or greater.

Voltage

Small servos were originally designed to run from 4.8V rechargeable batteries in model aircraft. They can be driven with 5VDC to 6VDC on a routine basis. A few servos are designed for higher voltages.

Amperage

The datasheets provided by most manufacturers often fail to specify the power that a servo will draw when it is exerting maximum torque (or indeed, any torque at all). Since small servos are often driven by three or four AA alkaline batteries in series, the maximum current draw is unlikely to be much greater than 1 amp. When the motor is energized but not turning, and is not resisting a turning force, its power consumption is negligible. This feature makes servos especially desirable for remote-controlled battery-powered devices.

Some motors that have a turning range exceeding 180 degrees will respond to pulses of less than 1ms or greater than 2ms. A newly acquired motor should be tested with a microcontroller

that steps through a wide range of pulse durations, to determine the limits empirically. Pulses that are outside the motor's designed range will generally be ignored and will not cause damage.

The *turn rate* or *transit time* specified in a datasheet is the time a servo takes to rotate through 60 degrees, with no load on the output shaft. A high-torque servo generally achieves its greater turning force by using a higher reduction gear ratio, which tends to result in a longer transit time.

How to Use it

Typical applications for a small servo include rotating the flaps or rudder of a model aircraft, steering a model boat, model car, or wheeled robot, and turning robotic arms.

A servo generally has three wires, colored red (power supply), black or brown (ground), and orange, yellow, or white (for the pulse train from the controller). The ground wire to the motor must be common with the ground of the controller, and consequently a ceramic bypass **capacitor** of 0.1 μ F or 0.01 μ F should be placed between the (red) power wire to the motor and ground. A protection **diode** should also be used. Neither a diode nor a capacitor should be attached to the wire carrying control signals, as it will interfere with the pulse train.

When powering the motor, an AC adapter should only be used with some caution, as its power output may be inadequately smoothed. A voltage regulator is not necessary, but bypass **capacitors** are mandatory. Figure 24-7 shows two hypothetical schematics. The upper section of the figure shows a battery-driven system, possibly using four 1.2V NiMH rechargeable batteries. Since batteries do not generally create voltage spikes, no capacitors are used, but a diode is included to protect the microcontroller from EMF when the servo stops and starts. The lower section of the figure shows the additional precautions that may be necessary when using DC power from an **AC adapter**. The DC-DC **converter**,

which derives 6VDC for the motor requires smoothing capacitors (this should be specified in its datasheet), and so does the **voltage regulator**, which delivers regulated 5VDC power to the microcontroller. Once again, the protection diode is included. In both diagrams, the orange wire represents the control wire transmitting pulses to the servo motor.

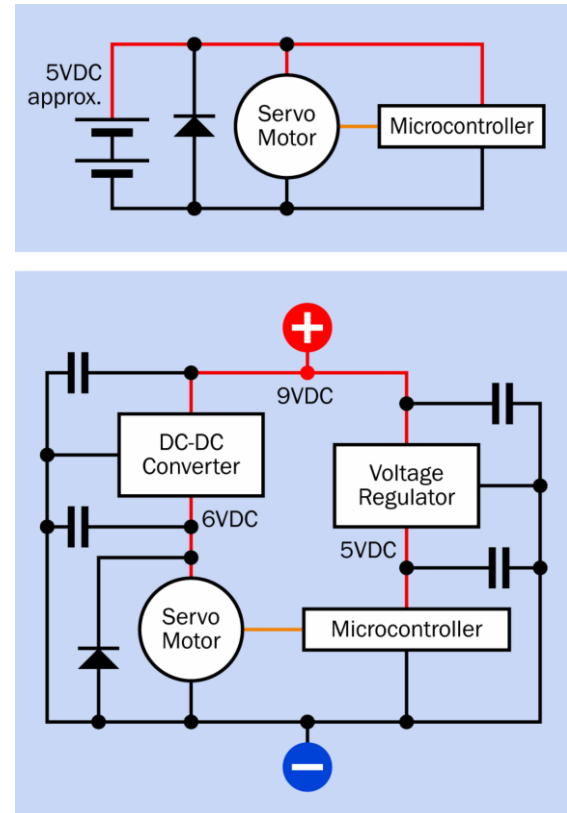


Figure 24-7. Two possible schematics to run a small servo motor, the upper example using battery power (for example, from four 1.2V NiMH cells) and the lower example using a 9VDC AC adapter. See text for additional explanation.

Various shaft attachments are available from the same online hobby-electronics suppliers that sell servos. The attachments include discs, single arms, double arms, and four arms in a cross-shaped configuration. A single-arm attachment is often known as a *horn*, and this term may be applied loosely to any kind of attachment. The

horn is usually perforated so that other components can be fixed to it by using small screws or nuts and bolts. [Figure 24-8](#) shows a variety of horns.



Figure 24-8. Various shaft attachments, known as horns, are available from motor manufacturers. The blue one is metallic; the others are plastic.

After the horn is pushed onto the spline (the motor shaft), it is held in place with one central screw. As previously noted, the two major manufacturers of small servos, Futaba and Hitec, have incompatible splines.

Modification for Continuous Rotation

It is possible to modify a small servo motor so that it will rotate continuously.

First the motor case must be opened, and the potentiometer must be centered by using a controller to send some 1.5ms pulses. The potentiometer must then be glued or otherwise secured with its wiper in this precise center position, after which the potentiometer is disconnected from the drive train.

Mechanical stops that would limit the rotation of the motor shaft must be cut away, after which the motor is reassembled. Because the potentiometer has been immobilized, the motor's internal electronics will now "see" the shaft as being in its center position at all times. If the controller sends a pulse instructing the motor to seek a position clockwise or counter-clockwise from center, the motor will rotate in an effort to reach that

position. Because the potentiometer will not provide feedback to signal that the motor has achieved its goal, the shaft will continue to rotate indefinitely.

In this mode, the primary distinguishing characteristic of the servo has been disabled, in that it can no longer turn to a specific angle. Also, stopping the servo may be problematic, as it must receive a command that precisely matches the fixed position of the potentiometer. Since the potentiometer may have moved fractionally during the process in which it was immobilized, some trial and error may be needed to determine the pulse width that corresponds with the potentiometer position.

The purpose of modifying a servo for continuous rotation is to take advantage of its high torque, small size, light weight, and the ease of controlling it with a microcontroller.

In response to the interest shown by hobbyists in modifying servos for continuous rotation, some manufacturers now market servos with continuous rotation as a built-in feature. Typically they include a trimmer potentiometer to calibrate the motor, to establish its center-off position.

What Can Go Wrong

Incorrect Wiring

The manufacturer's datasheet should be checked to confirm the color coding of the wires. While a simple DC motor can be reversed by inverting the polarity of its power supply, this is totally inappropriate for a servo motor.

Shaft/Horn Mismatch

Attachments for the spline of one brand of motor may not fit the spline of another brand, and cannot be forced to fit.

Unrealistically Rapid Software Commands

Microcontroller software that positions a servo must allow sufficient time for the servo to respond before the software specifies a new position. It may be necessary to insert delay loops or other wait times in the software.

Jitter

A servo arm that twitches unpredictably usually indicates that the pulse train is being corrupted by external electrical noise. The control wire to the servo should be as short as possible, and should not run closely adjacent to conductors carrying AC or high frequency current switching, or control wires for other servo motors.

Motor Overload

A servo capable of delivering 2 lbs of force 1 inch from its shaft can easily generate enough torque, when it stalls, to break itself free from its mounts, or bend or break any arm or linkage attached to its shaft. Ideally, a relatively “weak link” should be included so that if breakage occurs, it will be predictable and will be relatively easy and cheap to repair.

Unrealistic Duty Cycle

Small servos are designed for intermittent use. Constant cycling will cause wear and tear, especially if the motor has a brushed commutator or metal reduction gears.

Electrical Noise

Brushed motors are always a source of electrical interference, and any servo will also tend to create a voltage dip or surge when it starts and stops. A protection diode may be insufficient to isolate sensitive microcontrollers and other integrated circuit chips. To minimize problems, the servo can be driven by a source of positive voltage that is separate from the regulated power supply used by the chips, and larger filter capacitors may be added to the voltage supply of the microcontroller. A common ground between the motor and the chips is unfortunately unavoidable.

stepper motor

Also often referred to as a *stepping motor*, and sometimes known as a *step motor*. It is a type of *induction motor* but merits its own entry in this encyclopedia as it has acquired significant and unique importance in electronics equipment where precise positioning of a moving part is needed and digital control is available.

OTHER RELATED COMPONENTS

- **DC motor** (See [Chapter 22](#))
- **AC motor** (See [Chapter 23](#))
- **servo motor** (See [Chapter 24](#))

What It Does

A stepper motor rotates its drive shaft in precise steps in response to a timed sequence of pulses (usually one step per pulse). The pulses are delivered to a series of coils or *windings* in the *stator*, which is the stationary section of the motor, usually forming a ring around the *rotor*, which is the part of the motor that rotates. Steps may also be referred to as *phases*, and a motor that rotates in small steps may be referred to as having a high *phase count*.

A stepper motor theoretically draws power for its stator coils at a constant level that does not vary with speed. Consequently the torque tends to decrease as the speed increases, and conversely, it is greatest when the motor is stationary or locked.

The motor requires a suitable control system to provide the sequence of pulses. The control system may consist of a small dedicated circuit, or a microcontroller or computer with the addition of suitable driver transistors capable of handling

the necessary current. The torque curve of a motor can be extended by using a controller that increases the voltage as the speed of the control pulses increases.

Because the behavior of the motor is controlled by external electronics, and its interior is usually symmetrical, a stepper motor can be driven backward and forward with equal torque, and can also be held in a stationary position, although the stator coils will continue to consume power in this mode.

How It Works

The stator has multiple poles made from soft iron or other magnetic material. Each pole is either energized by its own coil, or more commonly, several poles share a single, large coil. In all types of stepper motor, sets of stator poles are magnetized sequentially to turn the rotor and can remain energized in one configuration to hold the rotor stationary.

The rotor may contain one or more permanent magnets, which interact with the magnetic fields

generated in the stator. Note that this is different from a [squirrel-cage AC motor](#) in which a “cage” is embedded in the rotor and interacts with a rotating magnetic field, but does not consist of permanent magnets.

Three small stepper motors are shown in [Figure 25-1](#). Clockwise from the top-left, they are four-wire, five-wire, and six-wire types (this distinction is explained in the following section). The motor at top-left has a threaded shaft that can engage with a collar, so that as the motor shaft rotates counter-clockwise and clockwise, the collar will be moved down and up.

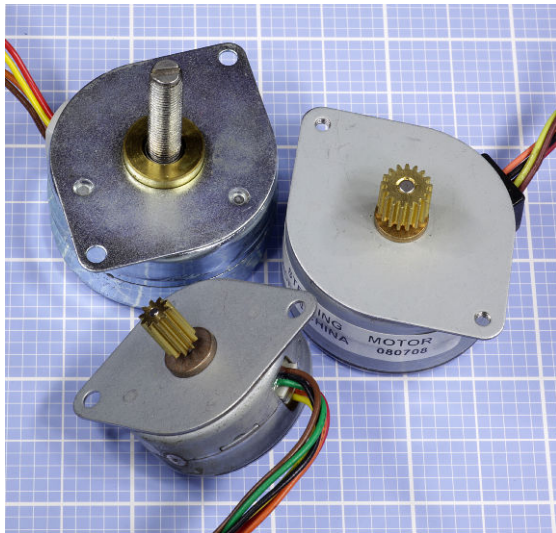


Figure 25-1. Three small stepper motors.

Reluctance Stepper Motors

The simplest form of stepper motor uses a rotor that does not contain permanent magnets. It relies on the principle of [variable reluctance](#), reluctance being the magnetic equivalent of electrical [resistance](#). The rotor will tend to align its protruding parts with the exterior source(s) of the magnetic field, as this will reduce the reluctance in the system. Additional information about variable reluctance is included in “[Reluctance Motor](#)” ([page 197](#)) in the section of this encyclopedia dealing with the **AC motor**.

A variable reluctance motor requires an external controller that simply energizes the stator coils sequentially. This is shown in [Figure 25-2](#), where six poles (energized in pairs) are arrayed symmetrically around a rotor with four protrusions, usually referred to as [teeth](#). Six stator poles and four teeth are the minimum numbers for reliable performance of a reluctance stepper motor.

In the diagram, the core of each pole is tinted green when it is magnetized, and is gray when it is not magnetized. In each section of this diagram, the stator coils are shown when they have just been energized, and the rotor has not yet had time to respond. External switching to energize the coils has been omitted for simplicity. In a real motor, the rotor would have numerous ridges, and the clearance between them and the stator would be extremely narrow to maximize the magnetic effect.

In a 6-pole reluctance motor where the rotor has four teeth, each time the controller energizes a new pair of poles, the rotor turns by 30 degrees counter-clockwise. This is known as the [step angle](#), and means that the motor makes 12 steps in each full 360-degree rotation of its shaft. This configuration is very similar to that of a 3-phase AC [induction motor](#), as shown in [Figure 23-11](#) in the **AC motor** section of this encyclopedia. However, the AC motor is designed to be plugged into a power source with a constant frequency, and is intended to run smoothly and continuously, not in discrete steps.

Generally, reluctance motors tend to be larger than those with magnetized rotors, and often require feedback from a sensor that monitors shaft angle and provides this information to control electronics. This is known as a [closed loop](#) system. Most smaller stepper motors operate in an [open loop](#) system, where positional feedback is considered unnecessary if the number of pulses to the motor is counted as a means of tracking its position.

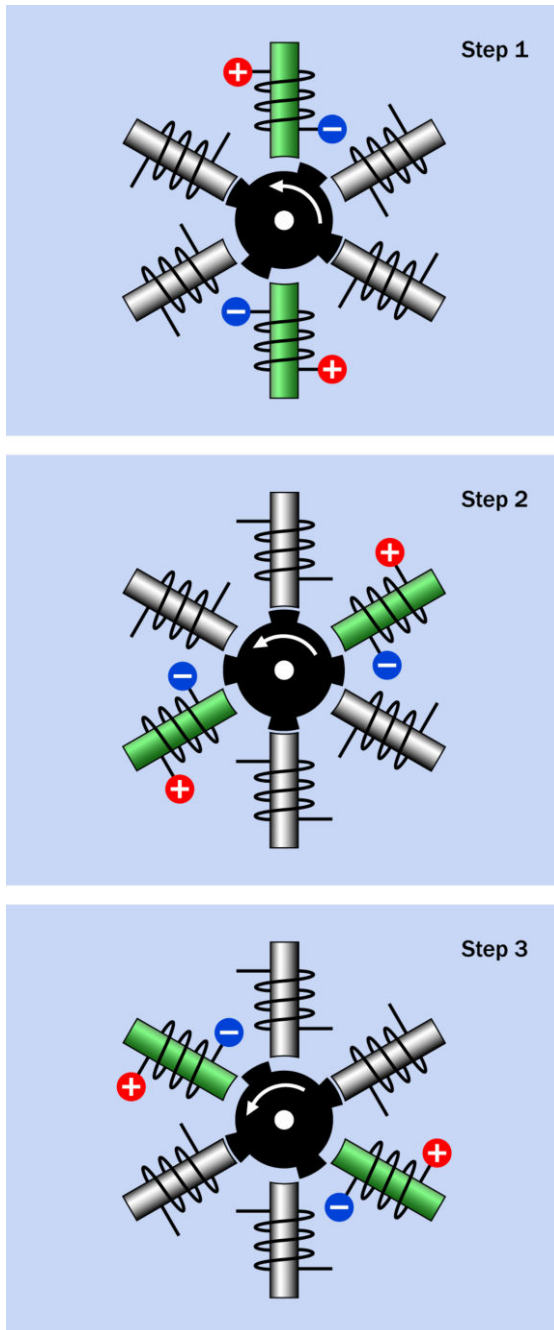


Figure 25-2. In a variable reluctance stepper motor, the rotor moves to minimize magnetic reluctance each time the next pair of coils is energized. At each step, the coils have been energized a moment before the rotor has had time to respond.

Permanent Magnet Stepper Motors

More commonly, the rotor of a stepper motor contains permanent magnets, which require the controller to be capable of reversing the magnetic field created by each of the stator coils, so that they alternately attract and repel the rotor magnets.

In a *bipolar* motor, the magnetic field generated by a coil is reversed simply by reversing the current through it. This is shown diagrammatically in Figure 25-3. In a *unipolar* motor, the magnetic field is reversed by applying positive voltage to the center tap of a coil, and grounding one end or the other. This is shown diagrammatically in Figure 25-4.

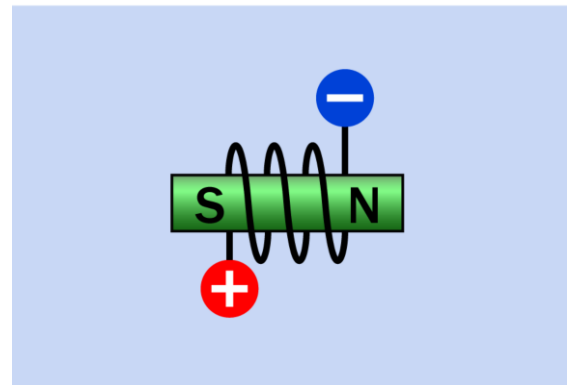
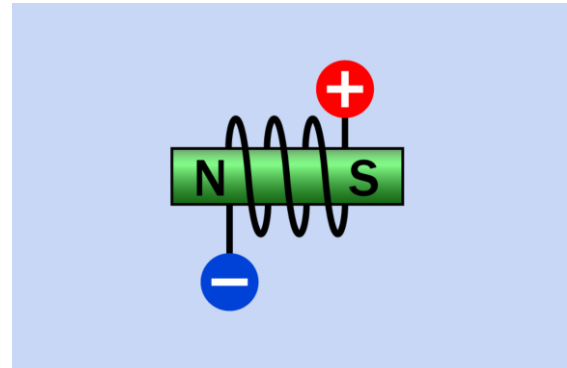


Figure 25-3. In a bipolar motor, the magnetic field generated by each stator coil is reversed simply by reversing the current through the coil.

Either type of motor is often designed with an upper and lower deck surrounding a single rotor,

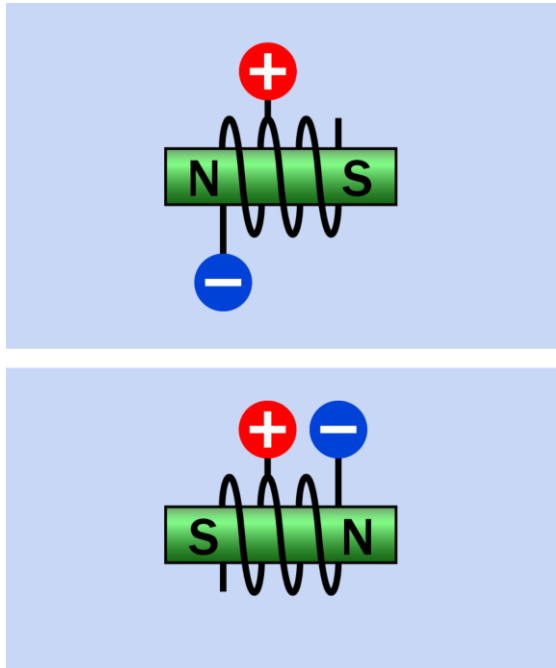


Figure 25-4. The magnetic field of this coil is reversed by applying positive voltage constantly to a center tap and grounding one end of the coil or the other.

as suggested in [Figure 25-5](#). A large single coil, or center-tapped coil, induces a magnetic field in multiple poles in the top deck, out of phase by one step with a second set of poles, energized by their own coil, in the bottom deck. (All three motors shown in [Figure 25-1](#) are of this type.) The rotor of the motor is tall enough to span both decks, and is rotated by each of them in turn.

In [Figure 25-6](#), the decks of a two-deck four-wire motor have been split apart. The rotor remains in the left-hand section. It is enclosed within a black cylinder that is a permanent magnet divided into multiple poles. In the right-hand section, a coil is visible surrounding metal “teeth” that function as stator poles when the coil is energized.

In [Figure 25-7](#), the same motor has been further disassembled. The coil was secured with a length of tape around its periphery, which has been removed to make the coil visible. The remaining

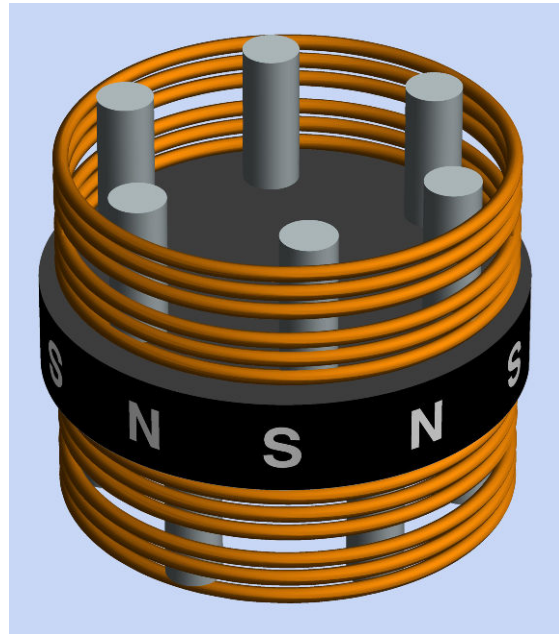


Figure 25-5. A simplified rendering of the common “two deck” type of motor. See text for details.

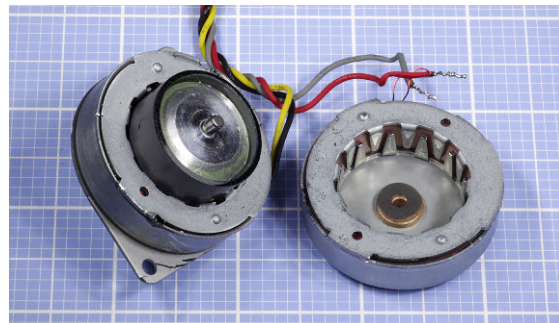


Figure 25-6. A two-deck stepper motor split open to reveal its rotor (left) and one of the stators (right) encircled by a coil.

half of the motor, at top-right, contains a second, concealed but identical coil with its own set of poles, one step out of phase with those in the first deck.

Because the field effects in a two-deck stepper motor are difficult to visualize, the remaining diagrams show simplified configurations with a minimum number of stator poles, each with its own coil.

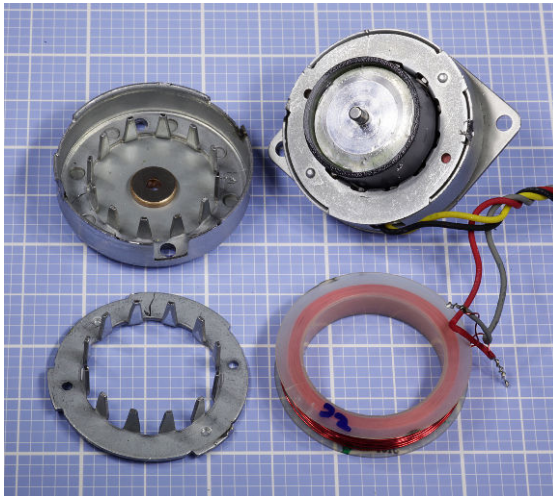


Figure 25-7. The stepper motor from the previous figure, further disassembled.

Bipolar Stepper Motors

The most basic way to reverse the current in a coil is by using an *H-bridge* configuration of switches, as shown in Figure 25-8, where the green arrow indicates the direction of the magnetic field. In actual applications, the switches are solid-state. Integrated circuits are available containing all the necessary components to control a bipolar stepper motor.

Four sequential steps of a bipolar motor are shown in Figure 25-9, Figure 25-10, Figure 25-11, and Figure 25-12. The H-bridge control electronics for each coil are omitted for clarity. As before, energized coils are shown with the pole inside the coil tinted green, while non-energized coils are gray, and the rotor is shown before it has had time to respond to the magnetic field in each step.

Unipolar Motors

The control electronics for a unipolar motor can be simpler than those for a bipolar motor, as off-the-shelf switching transistors can ground one end of the coil or the other. The classic five-wire unipolar stepper motor, often sold to hobbyists and used in robotics projects and similar applications, can be driven by nothing more elaborate

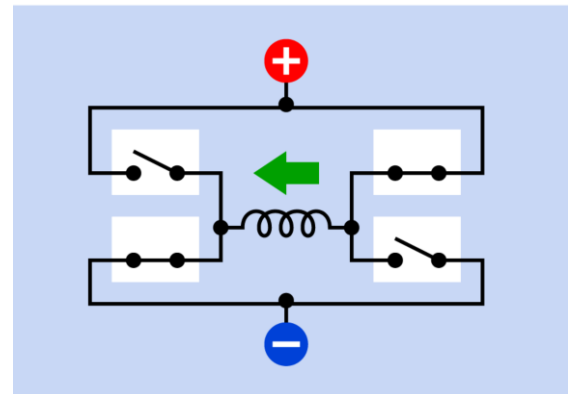
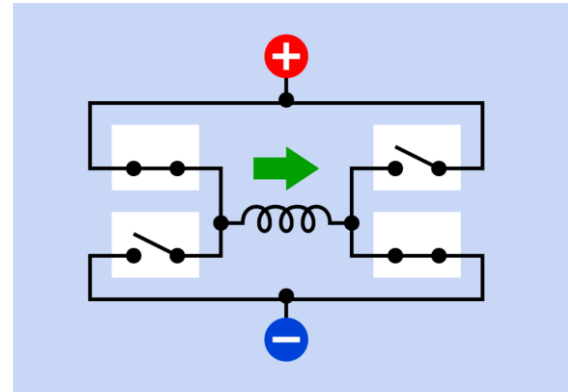


Figure 25-8. The simplest and most basic way to reverse the current through a coil is via an H-bridge circuit. In practice, the switches are replaced by solid-state components.

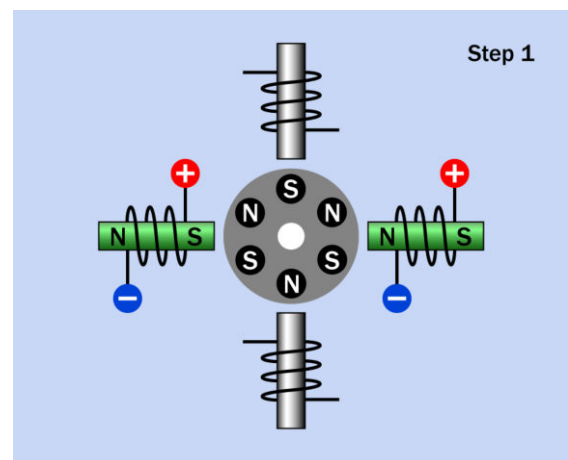


Figure 25-9. A bipolar stepper motor depicted a moment before the rotor has had time to make its first step in response to magnetic fields created by the stator coils.

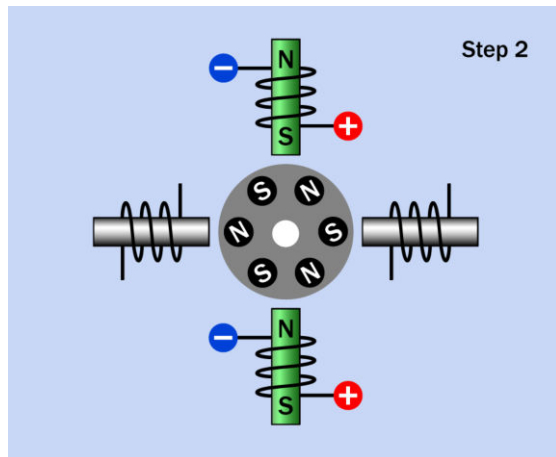


Figure 25-10. The bipolar stepper motor from the previous figure is shown with its rotor having advanced by one step, and coil polarity changed to induce it to make a second step.

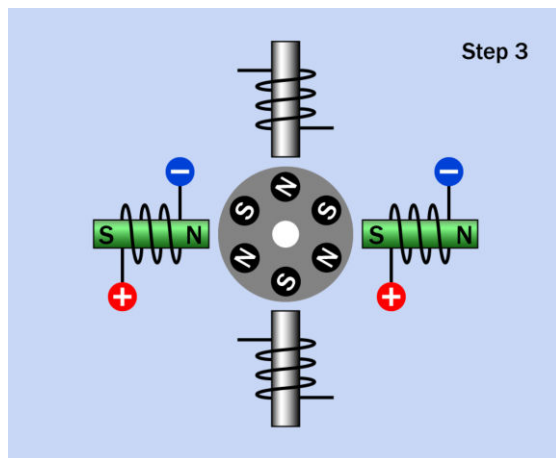


Figure 25-11. The bipolar stepper motor after taking its second step, immediately before making its third step.

than a set of *555 timer* chips. However, this type of motor is less powerful for its size and weight because only half of each coil is energized at a time.

In [Figure 25-13](#), [Figure 25-14](#), [Figure 25-15](#), and [Figure 25-16](#), the simplest configuration of a unipolar system is shown in diagrammatic form using four stator coils and a rotor containing six magnetic poles. Each figure shows the stator coils when they have just been energized, a mo-

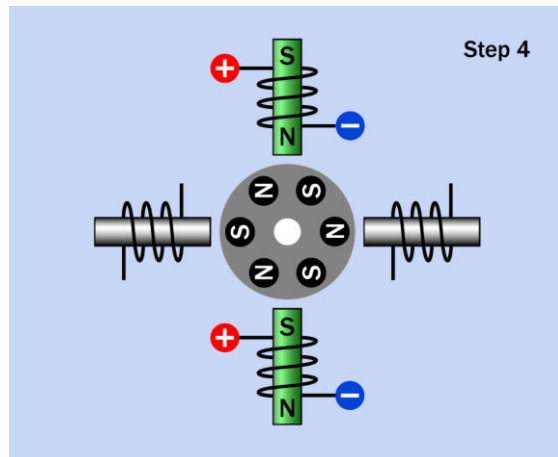


Figure 25-12. The bipolar stepper motor after taking its third step. When the rotor responds to the new pattern of magnetic fields, its orientation will be functionally identical with that shown in the first step.

ment before the rotor has had time to move in response to them. Coils that are energized are shown with the metal cores tinted green. Wires that are not conducting current are shown in gray. The open and closed positions of switches a, b, c, and d suggest the path that current is taking along the wires that are colored black.

Note that coils on opposite sides of the motor are energized simultaneously, while the other pair of coils is de-energized. Adjusting the controller so that it overlaps the “on” cycles of the coils can generate more torque, while consuming more power.

A motor containing more stator poles can advance in smaller steps, if the poles are separately energized. However, if the coils have individual windings, this will increase the cost of the motor.

Variants

In addition to bipolar and unipolar variants, previously described, three others are available.

High Phase Count

This term describes any type of stepper motor in which additional poles reduce the step size. The advantages of a high phase count include

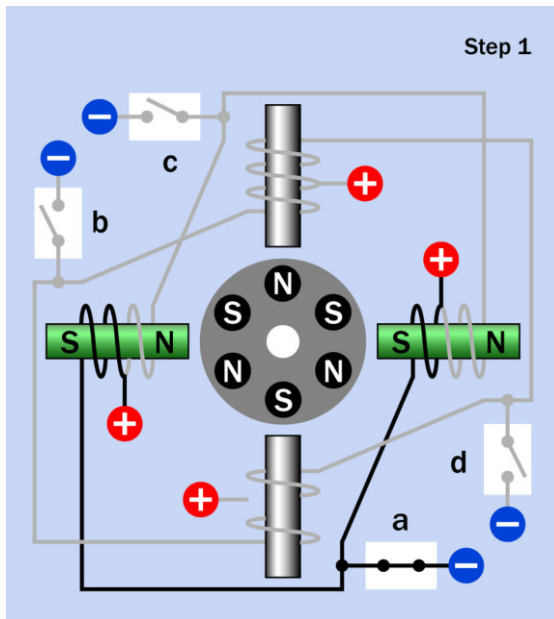


Figure 25-13. The coils of this unipolar stepper motor are shown an instant after they have been energized, before the rotor has had time to respond by making its first step.

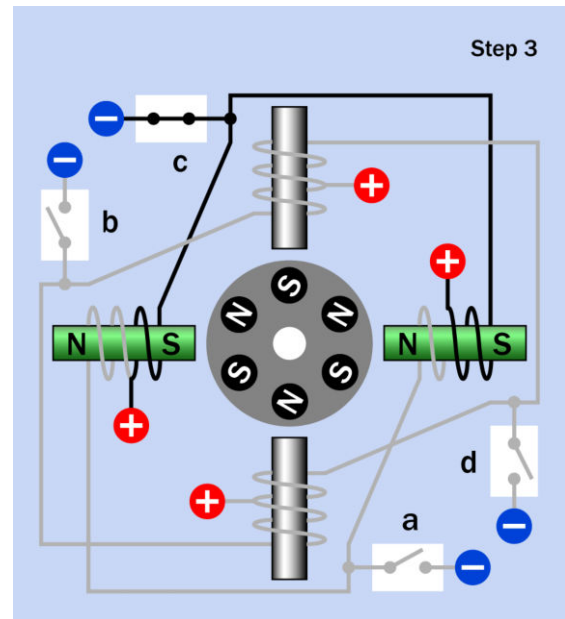


Figure 25-15. The same motor from the previous figure is shown with coils energized to induce the rotor to make its third step.

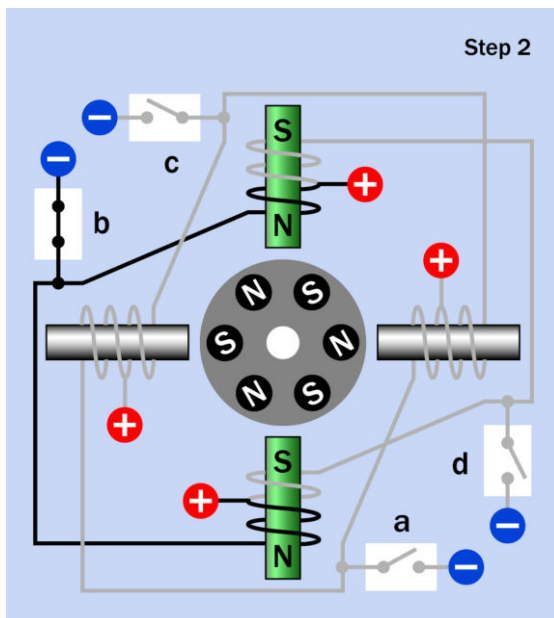


Figure 25-14. The same motor from the previous figure is shown with coils energized to induce the rotor to make its second step.

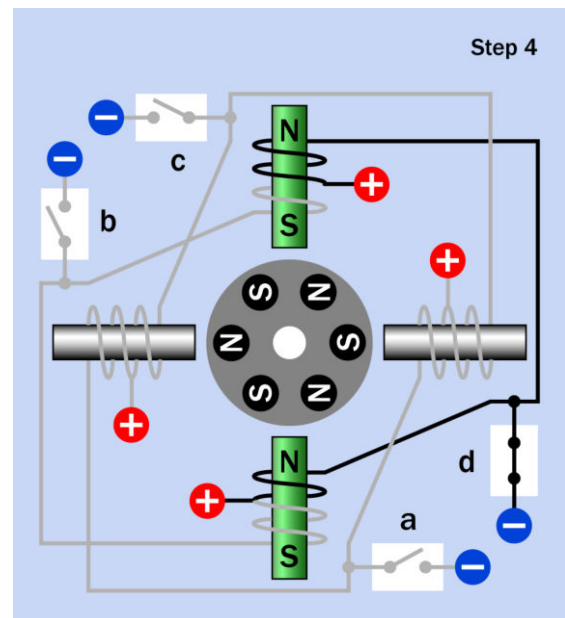


Figure 25-16. When the rotor makes its fourth step, it will be back in an orientation that is functionally identical with the first figure in this series.

smoother running at high speed and greater precision when selecting a desired motor position. The additional coils also enable higher power density, but naturally tend to add to the cost of the motor.

Hybrid

This type of motor uses a toothed rotor that provides variable reluctance while also containing permanent magnets. It has become relatively common, as the addition of teeth to the rotor enables greater precision and efficiency. From a control point of view, the motor behaves like a regular permanent-magnet stepper motor.

Bifilar

In this type of motor, also sometimes known as a *universal* stepper motor, two coils are wound in parallel for each stator pole. If there are two poles or sets of poles, and both ends of each winding are accessible via wires that are run out of the motor, there will be eight wires in total. Consequently this type is often referred to as an *8-wire motor*.

The advantage of this scheme is that it allows three possible configurations for the internal coils. By shorting together the wires selectively, the motor can be made to function either in unipolar or bipolar mode.

In [Figure 25-17](#), the upper pair of simplified diagrams depicts one end of one coil connected to the beginning of the other, while positive voltage is applied at the midpoint, as in a unipolar motor. The magnetic polarity of the coil is determined by grounding either end of the coil. The section of each coil that is not conducting current is shown in gray.

The center pair of diagrams shows the adjacent ends of the coils tied together, so that they are now energized in parallel, with the magnetic polarity being determined by the polarity of the voltage, as in a bipolar motor.

The coils may also be connected in series, as shown in the lower pair of diagrams. This will provide greater torque at low speed and lower torque at high speed, while enabling higher-voltage, lower-current operation.

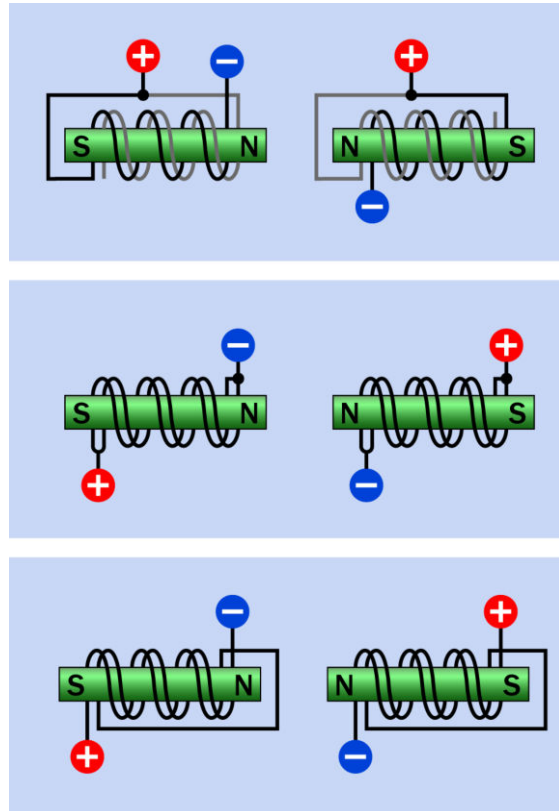


Figure 25-17. In a bifilar motor, two coils are wound in parallel around each stator pole and can be connected with a center tap to emulate a unipolar motor (upper diagrams), or can be energized in parallel (middle diagrams) or series (lower diagrams) to emulate a bipolar motor.

Multiphase

In a multiphase motor, multiple stator coils are usually connected in series, with a center tap applied between each pair. A possible configuration is shown in [Figure 25-18](#), where the two diagrams show two consecutive steps in rotation, although the step angle could be halved by changing the voltage polarity in only one location at a time. The way in which the motor is wired

enables only one stator coil to be unpowered during any step, because its two ends are at equal potential. Therefore this type of motor is capable of high torque in a relatively small format.

In some multiphase motors, additional wires allow access to both ends of each coil, and the coils are not connected internally. This allows control of the motor to be customized.

Microstepping

An appropriately designed stepper motor can be induced to make very small, intermediate steps if the control voltage is modulated to intermediate levels. Step angles as low as 0.007 degrees are claimed by some manufacturers. However, a motor running in this mode is less able to generate torque.

The simplest form of microstepping is half-stepping. To achieve this in a unipolar motor, each coil passes through an “off” state before its magnetic polarity is reversed.

Sensing and Feedback

So long as the series of pulses to the motor allows the rotor ample time to respond, no feedback mechanism from the rotor is necessary to confirm its position, and an open-loop system is sufficient. If sudden acceleration, deceleration, load fluctuations, and/or rotation reversal will occur, or if high speeds are involved, a *closed loop system*, in which a sensor provides positional feedback, may be necessary.

Voltage Control

Rapid stepping of a motor requires rapid creation and collapse of magnetic fields in the stator windings. Therefore, self-inductance of the windings can limit the motor speed. One way to overcome this is to use a higher voltage. A more sophisticated solution is to use a controller that provides a high initial voltage, which is reduced or briefly interrupted when a sensor indicates that coil current has increased sufficiently to overcome the self-inductance of the windings

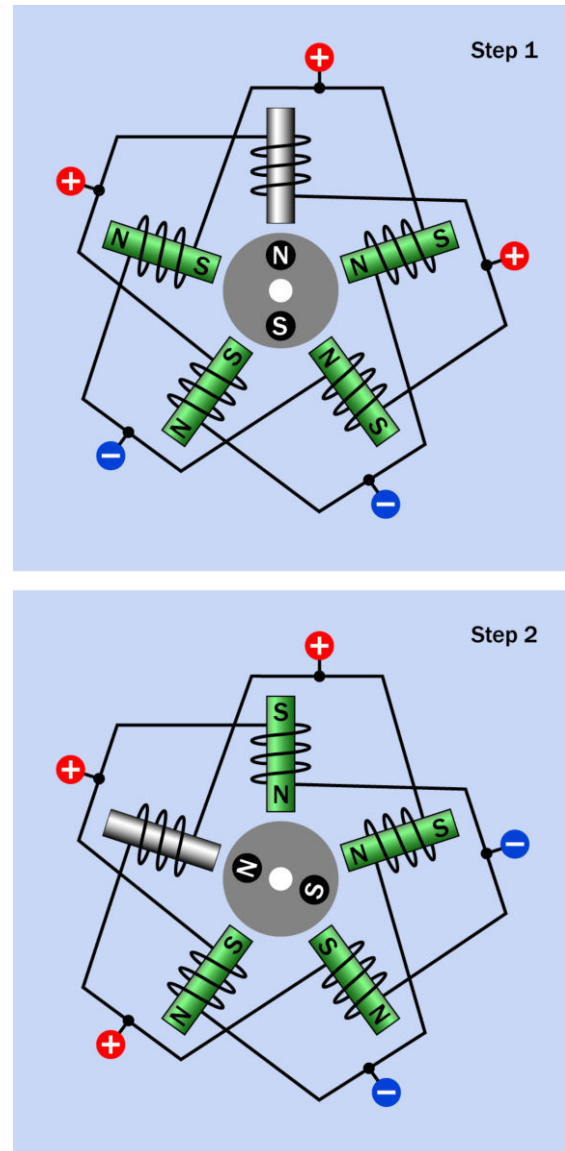


Figure 25-18. A multiphase stepper motor. By applying voltage in the pattern shown, only one coil is not energized during each step. This enables high torque compared with the size of the motor.

and has reached its imposed limit. This type of controller may be referred to as a *chopper drive* as the voltage is “chopped,” usually by power transistors. It is a form of *pulse width modulation*.

Values

The *step angle* of a stepper motor is the angular rotation of its shaft, in degrees, for each full step. This will be determined by the physical construction of the motor. The coarsest step angle is 90 degrees, while sophisticated motors may be capable of 1.8 degrees (without microstepping).

The maximum torque that a motor can deliver is discussed in “[Values](#)” (page 184) in the **DC motor** entry of this encyclopedia.

Motor weight and size, shaft length, and shaft diameter are the principal passive values of a stepper motor, which should be checked before it is selected for use.

How to Use it

Stepper motors are used to control the seek action in disk drives, the print-head movement and paper advance in computer printers, and the scanning motion in document scanners and copiers.

Industrial and laboratory applications include the adjustment of optical devices (modern telescopes are often oriented with stepper motors), and valve control in fluid systems.

A stepper motor may be used to power a *linear actuator*, usually via a screw thread (properly known as a lead screw) or worm gear. For more on linear actuators, see “[Linear Actuator](#)” (page 184). While the stepper motor will enable greater accuracy than a traditional DC motor, the gearing inevitably will introduce some imprecision.

Advantages of stepper motors include:

- Precise positioning, typically within 3 percent to 5 percent per step. The percentage step error does not accumulate as the motor rotates
- Able to run at a wide range of speeds, including very slow speeds without reduction gearing
- Trouble-free start, stop, and reverse action

- Cheap controller hardware where open-loop applications are acceptable
- High reliability, since no brushes or commutator are involved

Disadvantages include:

- Noise and vibration
- Resonance at low speeds
- Progressive loss of torque at high speeds

Protection Diodes

While a small stepper motor may be driven directly from power transistors, *darlington pairs*, or even *555 timers*, larger motors will create *back-EMF* when the magnetic field of each stator coil is induced or forward EMF when the field is allowed to collapse, and bipolar motors will also induce voltage spikes when the current reverses. In a unipolar motor, while only one-half of the coil is actually energized via its center tap, the other half will have an induced voltage, as the coil acts like a *linear transformer*.

A simplified schematic illustrating diode placement for a bipolar motor is shown in [Figure 25-19](#).

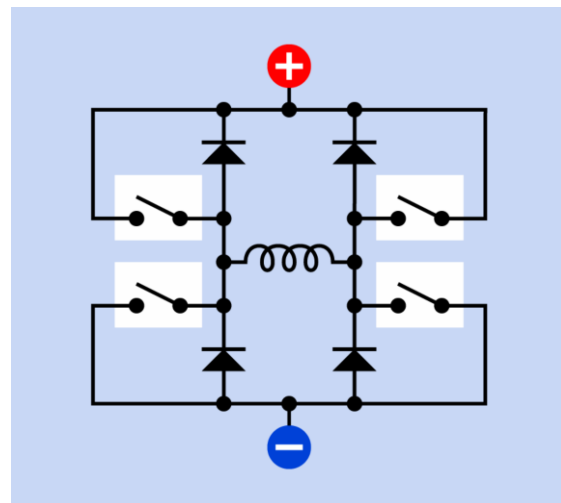


Figure 25-19. The H-bridge circuit must be augmented with protection diodes to guard against back-EMF created by fluctuating current in the stator coil.

Integrated circuit chips are available to incorporate protection diodes, in addition to the necessary power transistors. Stepper motors may also have protection diodes built in. Consult the manufacturer's datasheet for details before attaching a motor to a power source.

Positional Control

The built-in control electronics of a **servo motor** typically turn the shaft to a precisely known position in response to pulse-width modulation from an exterior source such as a **microcontroller**, whereas the angle of rotation of a stepper motor in an open-loop system must be calculated by counting the number of steps from an initial, home position. This limitation of a stepper motor can be overcome by using a closed-loop system, but that will require monitoring the motor, adding complexity to the external controller. The choice between stepper and servo motors should be evaluated on a case-by-case basis.

What Can Go Wrong

General problems affecting all types of motors are listed in [“Heat effects” \(page 188\)](#). Issues relating more specifically to stepper motors are listed in the following sections.

Incorrect Wiring

Because a stepper motor is driven via multiple conductors, there is a significant risk of wiring errors, especially since many motors are not identified with part numbers. The first challenge, then, may be to determine what type of motor it is. When the motor is disconnected from any power, and the shaft is rotated with finger and thumb, a magnetized-rotor motor will not spin as freely as a reluctance motor, because the magnets in the rotor will provide intermittent turning resistance.

If a unipolar motor is relatively small and is fitted with five wires, almost certainly the motor contains two coils, each with a center tap, and their function can be determined by applying positive

voltage to the red wire and grounding each of the other wires in turn. Attaching a small piece of tape to the motor shaft will assist in viewing its orientation.

A multimeter set to measure ohms can also be useful in deducing the internal coil connections of the motor, since the end-to-end resistance of a coil should be approximately twice the resistance between the center tap and either end of the coil.

A multiphase motor may have five wires, but in this case, the resistance between any two non-adjacent wires will be 1.5 times the resistance between any two adjacent wires.

Step Loss

In an open-loop system, if the motor skips or misses pulses from the controller, the controller no longer has an accurate assessment of the shaft angle. This is known as [step loss](#). Since this can be caused by sudden changes in control frequency, the frequency should be increased (or decreased) gradually. This is known as [ramping](#) the motor speed. Stepper motors cannot respond instantly to changes in speed, because of inertia in the rotor or in the device that the motor is driving.

Where the motor turns one or more steps beyond its commanded stopping point, this is known as [overshoot](#).

Step loss may also occur if the motor continues turning after power has been interrupted (either intentionally or because of an external fault). In an open-loop system, the controller should be designed to reset the motor position when power is initiated.

Excessive Torque

When the motor is stationary and not powered, [detent torque](#) is the maximum turning force that can be applied without causing the shaft to turn. When the motor is stationary and the controller does deliver power to it, [holding torque](#) is the maximum turning force that can be applied

without causing the shaft to turn, and *pull-in torque* is the maximum torque which the motor can apply to overcome resistance and reach full speed. When the motor is running, *pull-out torque* is the maximum torque the motor can deliver without suffering step loss (pulling it out of sync with its controller). Some or all of these values should be specified on the motor's datasheet. Exceeding any of them will result in step loss.

Hysteresis

When a controller directs a stepper motor to seek a specified position, the term *hysteresis* is often used to mean the total error between the actual position it reaches when turning clockwise, and the actual position it reaches when turning counter-clockwise. This difference may occur because a stepper motor tends to stop a fraction short of its intended position, especially under significant load. Any design that requires precision should be tested under real-world conditions to assess the hysteresis of the motor.

Resonance

A motor has a natural resonant frequency. If it is stepped near that frequency, vibration will tend to be amplified, which can cause positional errors, gear wear (if gears are attached), bearing wear, noise, and other issues. A good datasheet should specify the resonant frequency of the motor, and the motor should run above that frequency if possible. The problem can be addressed by rubber motor mounts or by using a resilient component, such as a drive belt, in conjunction with the drive shaft. *Damping* the vibration may be attempted by adding weight to the motor mount.

Note that if the motor has any significant weight attached directly to its shaft, this will lower its resonant frequency, and should be taken into account.

Resonance may also cause *step loss* (see preceding sections).

Hunting

In a closed-loop system, a sensor on the motor reports its rotational position to the controller, and if necessary, the controller responds by adjusting the position of the motor. Like any feedback system, this entails some lag time, and at certain speeds the motor may start *hunting* or *oscillating* as the controller over-corrects and must then correct its correction. Some closed-loop controllers avoid this issue by running mostly in open-loop mode, using correction only when the motor experiences conditions (such as sudden speed changes), which are likely to cause step loss.

Saturation

While it may be tempting to increase the torque from a stepper motor by upping the voltage (which will increase the current through the stator coils), in practice motors are usually designed so that the cores of the coils will be close to saturation at the rated voltage. Therefore, increasing the voltage may achieve very little increase in power, while causing a significant increase in heat.

Rotor Demagnetization

The permanent magnets in a rotor can be partially demagnetized by excessive heat. Demagnetization can also occur if the magnets are exposed to high-frequency alternating current when the rotor is stationary. Therefore, attempting to run a stepper motor at high speed when the rotor is stalled can cause irrevocable loss of performance.

diode

28

The term **diode** almost always means a semiconductor device, properly known as a *PN junction diode*, although the full term is not often used. It was formerly known as a *crystal diode*. Before that, **diode** usually meant a type of *vacuum tube*, which is now rarely used outside of high-wattage RF transmitters and some high-end audio equipment.

OTHER RELATED COMPONENTS

- **rectifier** (See “Rectification” (page 227))
- **unijunction transistor** (See Chapter 27)
- **LED** (light-emitting diode) (Volume 2)

What It Does

A diode is a two-terminal device that allows current to flow in one direction, known as the *forward direction*, when the *anode* of the diode has a higher positive potential than the *cathode*. In this state, the diode is said to be *forward biased*. If the polarity of the voltage is reversed, the diode is now *reverse biased*, and it will attempt to block current flow, within its rated limits.

Diodes are often used as *rectifiers* to convert alternating current into direct current. They may also be used to suppress voltage spikes or protect components that would be vulnerable to reversed voltage, and they have specialized applications in high-frequency circuits.

A *Zener diode* can regulate voltage, a *varactor diode* can control a high-frequency oscillator, and *tunnel diodes*, *Gunn diodes*, and *PIN diodes* have high-frequency applications appropriate to their rapid switching capability. An **LED** (*light-emitting diode*) is a highly efficient light source, which is discussed in Volume 2 of this encyclopedia. A **photosensitive diode** will adjust its ability to pass current depending on the light that falls upon it, and is included as a *sensor* in Volume 3.

See Figure 26-1 for schematic symbols representing a generic diode.

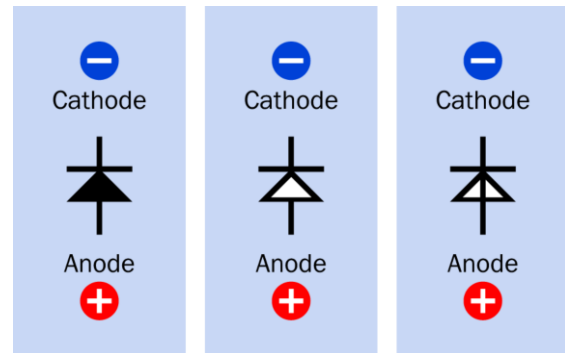


Figure 26-1. Commonly used schematic symbols for a generic diode. All the symbols are functionally identical. The direction of the arrow formed by the triangle indicates the direction of conventional current (from positive to negative) when the diode is forward-biased.

The basic diode symbol is modified in various ways to represent variants, as shown in Figure 26-2.

At top

Each symbol in the group of six indicates a Zener diode. All are functionally identical.

Bottom-left

Tunnel diode.

Bottom-center

Schottky diode.

Bottom-right

Varactor.

A triangle with an open center does not indicate any different function from a triangle with a solid center. The direction of the arrow always indicates the direction of conventional current, from positive to negative, when the diode is forward-biased, although the functionality of Zener diodes and varactors depends on them being reverse-biased, and thus they are used with current flowing opposite to the arrow symbol. The bent line used in the Zener symbol can be thought of as an opened letter Z, while the curled line used in the Schottky diode symbol can be thought of as a letter S, although these lines are sometimes drawn flipped left-to-right.

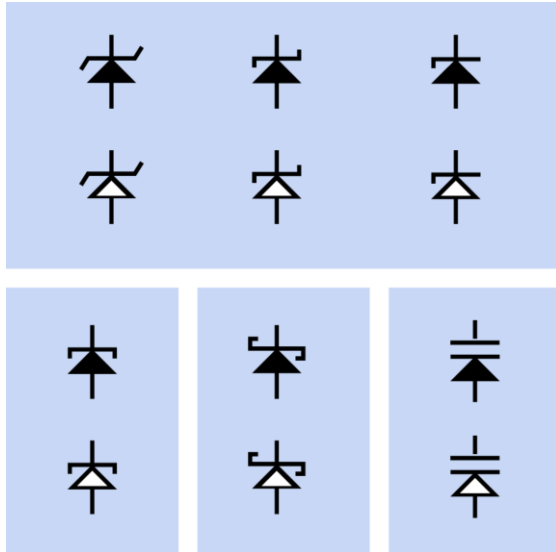


Figure 26-2. Commonly used schematic symbols for specialized types of diodes. See text for details.

A range of rectifier and signal diodes is shown in [Figure 26-3](#). (Top: Rectifier diode rated 7.5A at 35VDC. Second from top: Rectifier diode rated 5A at 35VDC. Center: Rectifier diode rated 3A at

35VDC. Second from bottom: 1N4001 Rectifier diode rated 1A at 35VDC. Bottom: 1N4148 signal switching diode rated at 300mA.) All values are for forward continuous current and RMS voltage. Each cylindrical diode is marked with a silver stripe (a black stripe on the 1N4148) to identify its *cathode*, or the end of the diode that should be “more negative” when the component is forward biased. Peak current can greatly exceed continuous current without damaging the component. Datasheets will provide additional information.

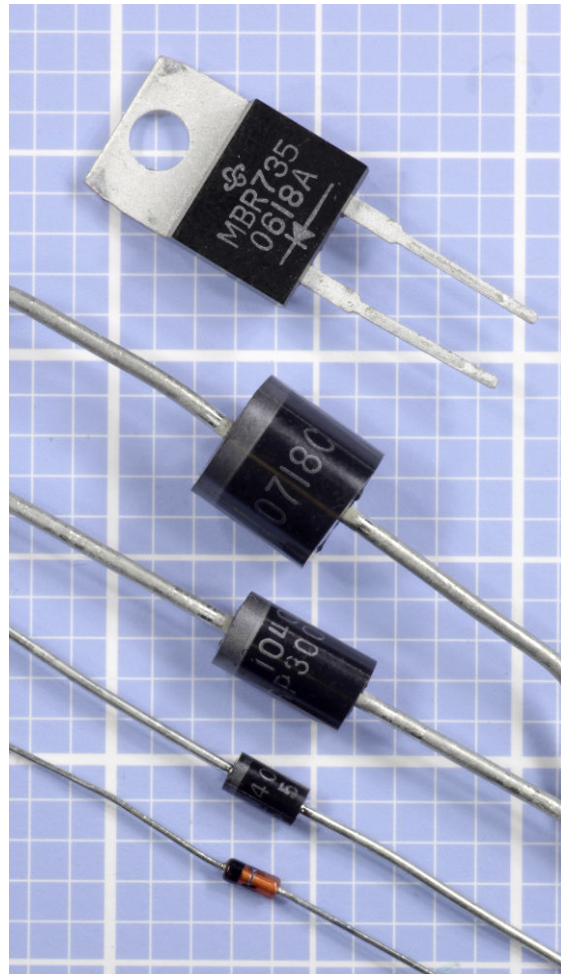


Figure 26-3. Diodes ranging in continuous forward-current capability from 7.5A (top) to 300mA (bottom). See text for additional details.

How It Works

A PN diode is a two-layer *semiconductor*, usually fabricated from silicon, sometimes from germanium, and rarely from other materials. The layers are doped with impurities to adjust their electrical characteristics (this concept is explained in more detail in [Chapter 28](#)). The *N layer* (on the negative, cathode side) has a surplus of electrons, creating a net negative charge. The *P layer* (on the positive, anode side) has a deficit of electrons, creating a net positive charge. The deficit of electrons can also be thought of as a surplus of “positive charges,” or more accurately, a surplus of *electron holes*, which can be considered as spaces that electrons can fill.

When the negative side of an external voltage source is connected with the cathode of a diode, and the positive side is connected with the anode, the diode is forward-biased, and electrons and electron holes are forced by mutual repulsion toward the junction between the n and p layers (see [Figure 26-4](#)). In a silicon diode, if the potential difference is greater than approximately 0.6 volts, this is known as the *junction threshold voltage*, and the charges start to pass through the junction. The threshold is only about 0.2 volts in a germanium diode, while in a Schottky diode it is about 0.4 volts.

If the negative side of an external voltage source is connected with the anode of a diode and positive side is connected with the cathode, the diode is now reverse-biased, and electrons and electron holes are attracted away from the junction between the n and p layers. The junction is now a *depletion region*, which blocks current.

Like any electronic component, a diode is not 100% efficient. When it is forward-biased and is passing current, it imposes a small voltage drop of around 0.7V for a silicon-based diode (Schottky diodes can impose a drop of as little as 0.2V, germanium diodes 0.3V, and some LEDs between 1.4V and 4V). This energy is dissipated as heat. When the diode is reverse-biased, it is still not 100% efficient, this time in its task of blocking

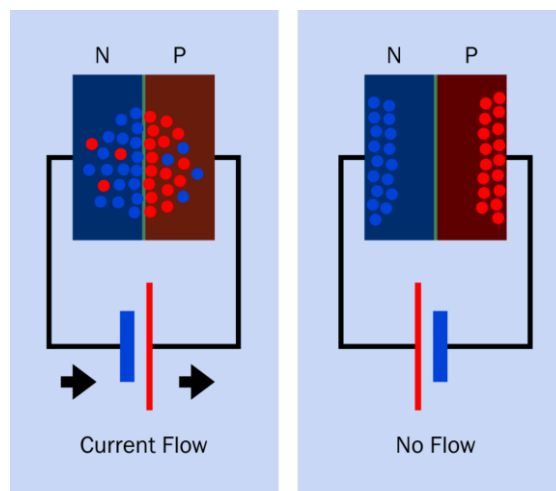


Figure 26-4. Inside a PN junction diode. Left: in forward-biased mode, voltage from a battery (bottom, with plates colored for clarity) forces charges in the N and P layers toward the central junction of the diode. Current begins to flow. Right: in reverse-biased mode, charges in the N and P layers are attracted away from the central junction, which becomes a depletion region, unable to pass significant current.

current. The very small amount of current that manages to get through is known as *leakage*. This is almost always less than 1mA and may be just a few μA , depending on the type of diode.

The performance of a theoretical generic PN diode is illustrated in [Figure 26-5](#). The right-hand side of the graph shows that if a diode is forward-biased with a gradually increasing potential, no current passes until the diode reaches its junction threshold voltage, after which the current rises very steeply, as the *dynamic resistance* of the diode diminishes to near zero. The left-hand side of the graph shows that when the diode is reverse-biased with a gradually increasing potential, initially a very small amount of current passes as leakage (the graph exaggerates this for clarity). Eventually, if the potential is high enough, the diode reaches its intrinsic *break-down voltage*, and once again its effective resistance diminishes to near zero. At either end of the curve, the diode will be easily and permanently

damaged by excessive current. With the exception of Zener diodes and varactors, reverse bias on a diode should not be allowed to reach the breakdown voltage level.

The graph in [Figure 26-5](#) does not have a consistent scale on its Y axis, and in many diodes the magnitude of the (reverse-biased) breakdown voltage will be as much as 100 times the magnitude of the (forward-biased) threshold voltage. The graph has been simplified for clarity.

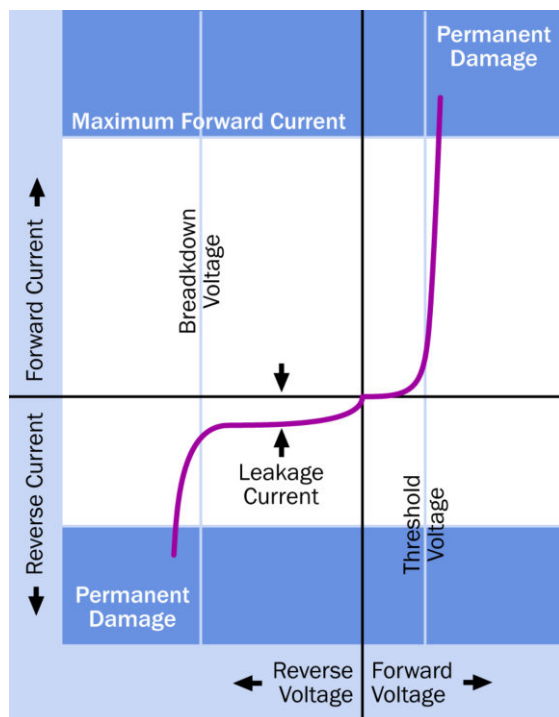


Figure 26-5. As the forward voltage across a diode reaches the junction threshold, the diode begins passing current. If the voltage across the diode is reversed, initially a small amount of current leakage occurs. Excessive forward or reverse voltage will create sufficient current to destroy the component.

Variants

Packaging

Some diodes have no information at all printed on them, while others may have a part number. Any additional information is rare. No convention exists for indicating the electrical character-

istics of the component by colors or abbreviations. If one terminal is marked in any way, almost certainly it is the cathode. One way to remember the meaning of a stripe on the cathode end of a rectifier diode or signal diode is by thinking of it as resembling the line in the diode schematic symbol.

Signal Diodes

Also known as *switching diodes* and *high-speed diodes*, their small size provides a low junction capacitance, enabling fast response times. They are not designed to withstand high currents. Signal diodes traditionally were packaged with axial leads for through-hole installation (like traditional-style resistors). Although this format still exists, signal diodes are now more commonly available in surface-mount formats.

Rectifier Diodes

Physically larger than signal diodes, and capable of handling higher currents. Their higher junction capacitance makes them unsuitable for fast switching. Rectifier diodes often have axial leads, although different package formats are used where higher currents are involved, and may include a heat sink, or may have provision for being attached to a heat sink.

There are no generally agreed maximum or minimum ratings to distinguish signal diodes from rectifier diodes.

Zener Diode

A Zener diode generally behaves very similarly to a signal or rectifier diode, except that its breakdown voltage is lower.

The Zener is intended to be reverse-biased; that is, conventional current is applied through it “in the wrong direction” compared with conventional diodes. As the current increases, the *dynamic resistance* of the Zener diode decreases. This relationship is shown in [Figure 26-6](#), where the two colored curves represent the performance of different possible Zener diodes. (The curves are adapted from a manufacturer’s data-

sheet.) This behavior allows the Zener to be used in simple voltage-regulator circuits, as it can allow a reverse current to flow at a voltage limited by the diode's breakdown voltage. Other applications for Zener diodes are described in “[DC Voltage Regulation and Noise Suppression](#)” (page 230). A typical Zener diode is shown in Figure 26-7.

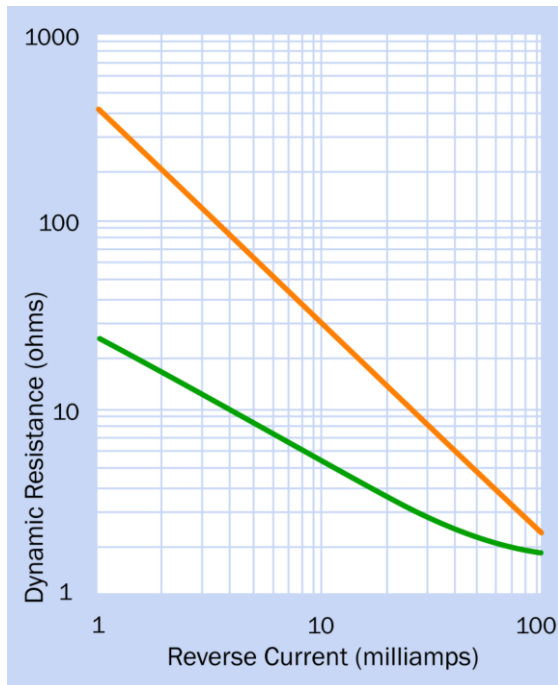


Figure 26-6. A manufacturer's datasheet may include graphs of this kind, showing the variation in dynamic resistance of two reverse-biased Zener diodes in response to changes in current.

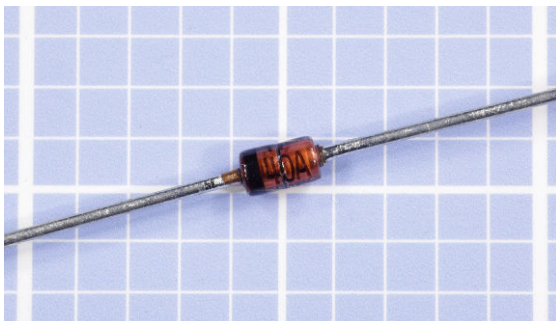


Figure 26-7. A 1N4740 Zener diode.

Transient Voltage Suppressor (TVS)

A form of Zener diode designed to protect sensitive devices from transient voltage spikes by clamping them—in other words, diverting the energy to ground. A TVS can absorb as much as 30,000 volts from a lightning strike or static discharge. Typically the Zener diode is incorporated in a network of other diodes in a surface-mount integrated circuit chip.

Zener diodes can also be used in circuits to handle electrostatic discharge (ESD), which can occur when a person unknowingly accumulates an electrostatic potential and then grounds it by touching an electronic device.

Schottky Diode

This type has a low junction capacitance, enabling faster switching than comparable generic silicon diodes. It also imposes a lower forward voltage drop, which can be desirable in low-voltage applications, and allows less power dissipation when a diode is necessary to control current flow. The Schottky diode is fabricated with a semiconductor-to-metal junction, and tends to be slightly more expensive than generic silicon diodes with similar voltage and current specifications.

Varactor Diode

Also known as a *varicap*, this type of diode has variable capacitance controlled by reverse voltage. While other diodes may exhibit this same phenomenon, the varactor is specifically designed to exploit it at very high frequencies. The voltage expands or contracts the depletion region in the junction between the P and N regions, which can be thought of as analogous to moving the plates of a capacitor nearer together or farther apart.

Because the capacitance of a varactor has a low maximum of about 100pF, its uses are limited. It is used extensively in RF applications where its voltage-controlled variable capacitance provides a uniquely useful way to control the

frequency of an oscillator circuit. In almost all radio, cellular, and wireless receivers, a varactor controls a phase-locked loop oscillator. In ham radio receivers, it can be used to adjust the tuning of a filter that tracks an incoming radio frequency.

A varactor is always reverse-biased below its breakdown voltage, so that there is no direct conduction. The voltage that controls a varactor must be absolutely free from random fluctuations that would affect its resonant frequency.

Tunnel Diode, Gunn Diode, PIN Diode

Mostly used in very high frequency or microwave applications, where ordinary diodes are unacceptable because they have insufficiently high switching speeds.

Diode Array

Two or more diodes may be encapsulated in a single DIP or (more commonly) surface-mount integrated circuit chip. The internal configuration and the pinouts of the chip will vary from one device to another. Diode arrays may be used for termination of data lines to reduce reflection noise.

Bridge Rectifier

Although this is a diode array, it is commonly indexed in parts catalogues under the term *bridge rectifier*. Numerous through-hole versions are available with ratings as high as 25A, some designed for single-phase input while others process three-phase AC. Screw-terminal components can rectify more than 1,000 volts at 1,000 amps. The package does not usually include any provision for smoothing or filtering the output. See “Rectification” (page 227) for more information on the behavior of a bridge rectifier.

Values

A manufacturer's datasheet for a typical generic diode should define the following values, using abbreviations that may include those in the following list.

- Maximum sustained forward current: I_f or I_o or I_{Omax}
- Forward voltage (the voltage drop imposed by the diode): V_f
- Peak inverse DC voltage (may be referred to as maximum blocking voltage or breakdown voltage): P_{iv} or V_{dc} or V_{br}
- Maximum reverse current (also referred to as leakage): I_r

Datasheets may include additional values when the diode is used with alternating current, and will also include information on peak forward surge current and acceptable operating temperatures.

A typical signal diode is the 1N4148 (included at the bottom of [Figure 26-3](#)), which is limited to about 300mA forward current while imposing a voltage drop of about 1V. The component can tolerate a 75V peak inverse voltage. These values may vary slightly among different manufacturers.

Rectifier diodes in the 1N4001/1N4002/1N4003 series have a maximum forward current of 1A and will impose a voltage drop of slightly more than 1V. They can withstand 50V to 1,000V of inverse voltage, depending on the component. Here again, the values may vary slightly among different manufacturers.

Zener diodes have a different specification, as they are used with reverse bias as voltage-regulating devices rather than rectification devices. Manufacturers' data sheets are likely to contain the following terminology:

- Zener voltage (the potential at which the diode begins to allow reverse current flow when it is reverse-biased, similar to breakdown voltage): V_z
- Zener impedance or dynamic resistance (the effective resistance of the diode, specified when it is reverse-biased at the Zener voltage): Z_z

- Maximum or admissible Zener current (or reverse current): I_z or I_{zm}
- Maximum or total power dissipation: P_d or P_{tot}

Zener voltage may be defined within a minimum and maximum range, or as a simple maximum value.

Limits on forward current are often not specified, as the component is not intended to be forward-biased.

How to Use it

Rectification

A *rectifier diode*, as its name implies, is commonly used to rectify alternating current—that is, to turn AC into DC. A half-wave rectifier uses a single diode to block one-half of the AC sinewave. The basic circuit for a half-wave rectifier is shown in [Figure 26-8](#). At top, the diode allows current to circulate counter-clockwise through the load. At bottom, the diode blocks current that attempts to circulate clockwise. Although the output has “gaps” between the pulses, it is usable for simple tasks such as lighting an LED, and with the addition of a smoothing capacitor, can power the coil of a DC relay.

A full-wave bridge rectifier employs four diodes to provide a more efficient output, usually filtered and smoothed with appropriate **capacitors**. The basic circuit is shown in [Figure 26-9](#). A comparison of input and output waveforms for half-wave and full-wave rectifiers appears in [Figure 26-10](#).

Discrete components are seldom used for this purpose, as off-the-shelf bridge rectifiers are available in a single integrated package. Rectifier diodes as discrete components are more likely to be used to suppress *back-EMF* pulses, as described below.

An old but widely used design for a full-wave bridge rectifier is shown in [Figure 26-11](#). This unit measured approximately 2" × 2" × 1.5" and was

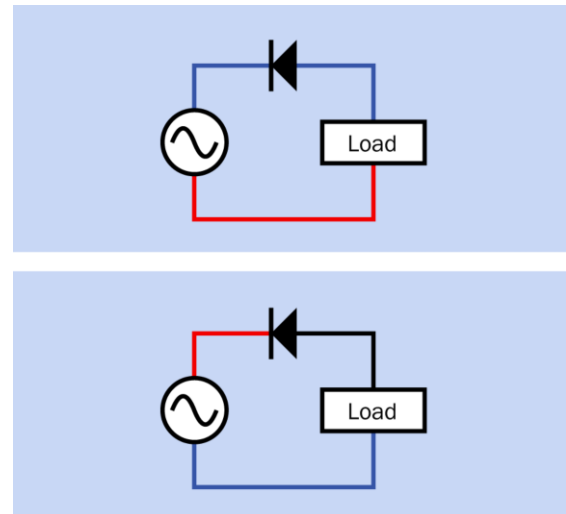


Figure 26-8. A half-wave rectifier. In this configuration the diode allows AC current to circulate counter-clockwise but blocks it clockwise.

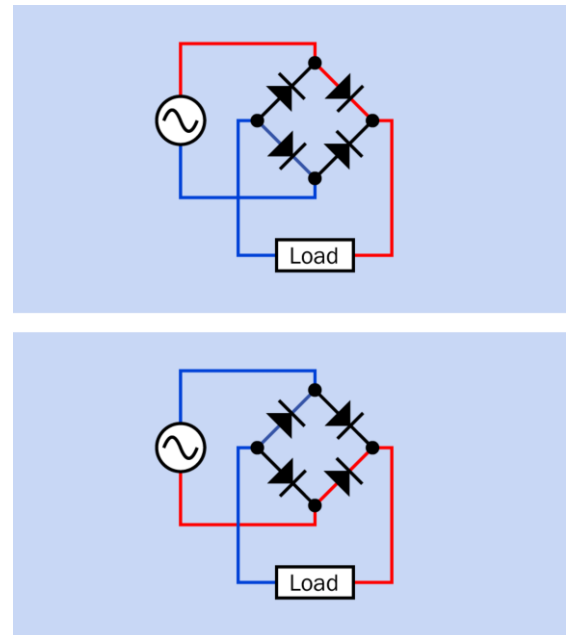


Figure 26-9. The basic circuit commonly used to form a bridge rectifier, with color added to indicate polarity. Wires shown in black are not passing current because diodes are blocking it. Note that the polarity at the load remains constant.

divided into four sections (as indicated by the solder terminals on the right-hand side), each

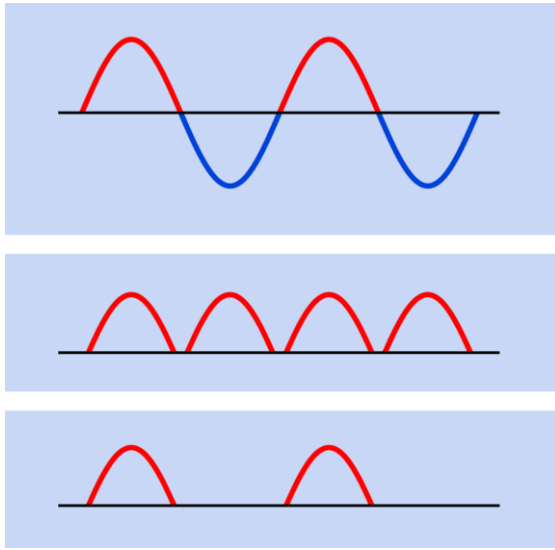


Figure 26-10. Top: The voltage-amplitude sinewave of an alternating current source that fluctuates between positive voltage (shown red) and negative voltage (shown blue) relative to a neutral (black) baseline. Center: AC current converted by a full-wave rectifier. Because the diodes do not conduct below their threshold voltage, small gaps appear between pulses. Bottom: Output from a half-wave rectifier.

section corresponding with the functionality of one modern diode. Figure 26-12 shows relatively modern rectifier packages, the one on the left rated at 20A continuous at 800V RMS, the one on the right rated 4A continuous at 200V RMS. In Figure 26-13, the one on the left is rated 4A continuous at 50V RMS, whereas the one on the right is rated 1.5A at 200V RMS.

DC output from rectifier packages is usually supplied via the outermost pins, while the two pins near the center receive AC current. The positive DC pin may be longer than the other three, and is usually marked with a + symbol.

Full-wave bridge rectifiers are also available in surface-mount format. The one in Figure 26-14 is rated for half an amp continuous current.

Back-EMF Suppression

A relay coil, motor, or other device with significant inductance will typically create a spike of voltage when it is turned on or off. This *EMF* can

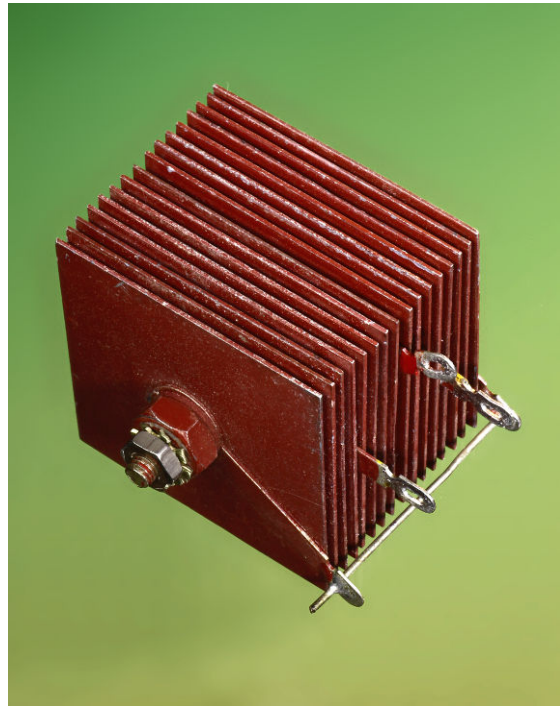


Figure 26-11. Prior to the perfection of chip fabrication in the late 1960s, it was common to find silicon rectifiers of this type, measuring about 2" square.

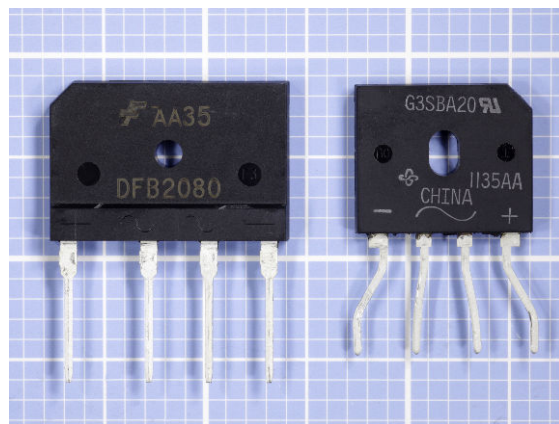


Figure 26-12. Full-wave bridge rectifies are commonly available in packages such as these. See text for details.

be shunted through a rectifier diode to safeguard other components in the circuit. A diode in this configuration may be referred to as a *protection diode*, a *clamp diode*, or *transient suppressor*. See Figure 26-15.

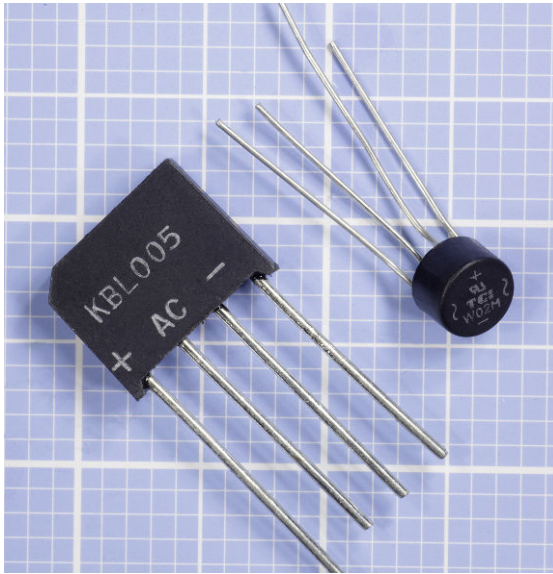


Figure 26-13. Smaller full-wave bridge rectifiers capable of 1.5A to 4A continuous current.



Figure 26-14. This surface-mount component contains four diodes forming a full-wave bridge rectifier circuit, and can pass 0.5A continuous current. It measures approximately 0.2" square.

Voltage Selection

A diode is sensitive to the relative voltage between its anode and cathode terminals. In other words, if the cathode is at 9V relative to the ground in the circuit, and the anode is at 12V, the 3V difference will easily exceed the threshold voltage, and the diode will pass current. (Actual tolerable values will depend on the forward voltage capability of the diode.) If the voltages are reversed, the diode will block the current.

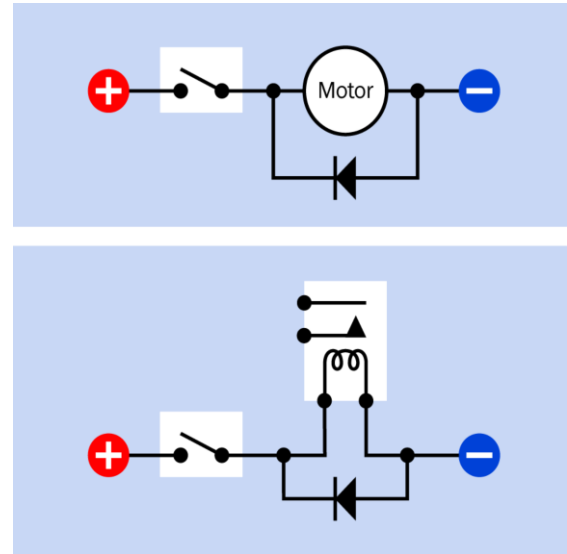


Figure 26-15. A rectifier diode is very often placed across a motor (top), relay (bottom), or other device with significant inductance that creates a spike of reverse voltage when switched on or off. The surge is shunted through the diode, protecting other components in the circuit.

This attribute can be used to make a device choose automatically between an AC adapter and a 9V battery. The schematic is shown in Figure 26-16. When an AC adapter that delivers 12VDC is plugged into a wall outlet, the adapter competes with the battery to provide power to a voltage regulator. The battery delivers 9VDC through the lower diode to the cathode side of the upper diode, but the AC adapter trumps it with 12VDC through the upper diode. Consequently, the battery ceases to power the circuit until the AC adapter is unplugged, at which point the battery takes over, and the upper diode now prevents the battery from trying to pass any current back through the AC adapter.

The voltage regulator in this schematic accepts either 12VDC or 9VDC and converts it to 5VDC. (In the case of 12VDC, the regulator will waste more power, which will be dissipated as waste heat.)

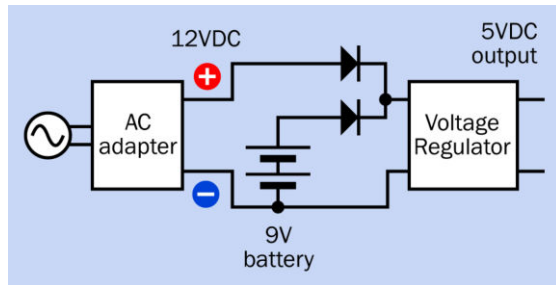


Figure 26-16. Two diodes with their cathodes tied together will choose automatically between an AC adapter that delivers 12VDC and an internal 9V battery.

Voltage Clamping

A diode can be used to *clamp* a voltage to a desired value. If an input to a 5V CMOS semiconductor or similarly sensitive device must be prevented from rising out of range, the anode of a diode can be connected to the input and the cathode to a 5V voltage source. If the input rises much above 5.6V, the potential difference exceeds the diode's junction threshold, and the diode diverts the excess energy. See Figure 26-17.

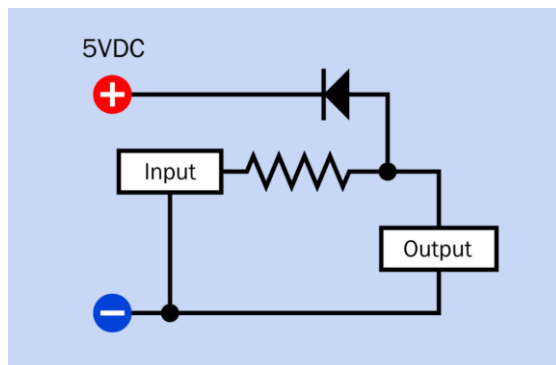


Figure 26-17. A clamping diode can limit output voltage—in this example, to about 5.6V. If the input rises above that value relative to the common ground, the potential difference across the diode feeds the excess voltage back through it to the 5V source.

Logic Gate

A signal diode is less than ideal as a logic gate, because it imposes a typical 0.6V voltage reduction, which can be significant in a 5V circuit and is probably unacceptable in a 3.3V circuit. Still, it can be useful on the output side—for example,

if two or more outputs from a logic chip or microcontroller are intended to drive, or share, another device such as a single **LED**, as shown in Figure 26-18. In this role, the diodes wired in parallel behave similarly to an OR gate, while preventing either output from the chip from feeding current back into the other output.

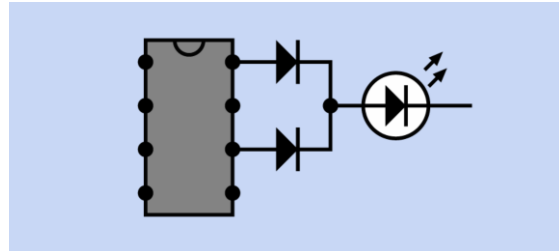


Figure 26-18. Two or more outputs from a logic chip or microcontroller may be coupled with diodes to power another device, such as an LED, while protecting the chip from backflow of current. The diodes form a logical OR gate.

DC Voltage Regulation and Noise Suppression

As previously noted, the dynamic resistance of a reverse-biased Zener diode will diminish as the current increases. This relationship begins at the point where breakdown in the diode begins—at its *Zener voltage*--and is approximately linear over a limited range.

The unique behavior of the Zener makes it usable as a very simple voltage controller when placed in series with a resistor as shown in Figure 26-19. It is helpful to imagine the diode and the resistor as forming a kind of voltage divider, with power being taken out at point A in the schematic. If a supply fluctuation increases the input voltage, this will tend to increase the current flowing through the Zener, and its dynamic resistance will diminish accordingly. A lower resistance in its position in the voltage divider will reduce the output voltage at point A, thus tending to compensate for the surge in input voltage.

Conversely, if the load in the circuit increases, and tends to pull down the input voltage, the current

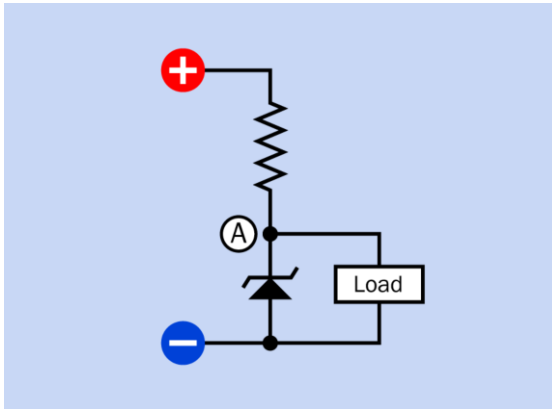


Figure 26-19. A simplified, basic circuit illustrating the ability of a Zener diode to compensate for variations in the power supply or load in a circuit, creating an approximately constant voltage at point A.

flowing through the Zener will diminish, and the voltage at point A will tend to increase, once again compensating for the fluctuation in the circuit.

As the series resistor would be a source of heat, a transistor could be added to drive the load, as shown in Figure 26-20.

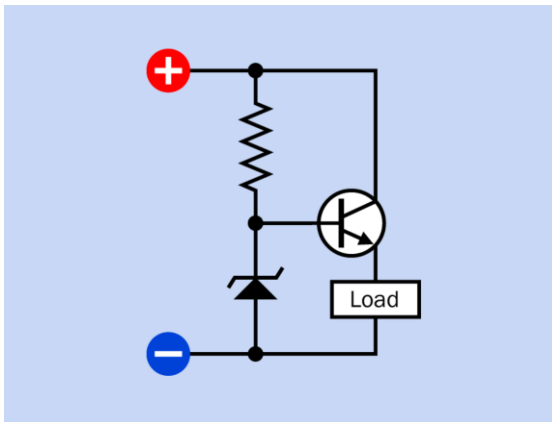


Figure 26-20. A transistor could be added to the circuit in the previous figure to reduce power waste through the resistor.

A manufacturer's datasheet may provide guidance regarding the dynamic resistance of a Zener diode in response to current, as previously shown in Figure 26-6. In practice, a packaged

voltage regulator such as the LM7805 would most likely be used instead of discrete components, since it includes self-calibrating features, requires no series resistor, and is relatively unaffected by temperature. However, the LM7805 contains its own Zener diode, and the principle of operation is still the same.

AC Voltage Control and Signal Clipping

A more practical Zener application would be to limit AC voltage and/or impose clipping on an AC sinewave, using two diodes wired in series with opposed polarities. The basic schematic is shown in Figure 26-21, while clipping of the AC sinewave is illustrated in Figure 26-22. In this application, when one diode is reverse-biased, the other is forward-biased. A forward-biased Zener diode works like any other diode: it allows current to pass relatively freely, so long as the voltage exceeds its threshold. When the AC current reverses, the Zeners trade their functions, so that the first one merely passes current while the second one limits the voltage. Thus, the diodes divert peak voltage away from the load. The Zener voltage of each diode would be chosen to be a small margin above the AC voltage for voltage control, and below the AC voltage for signal clipping.

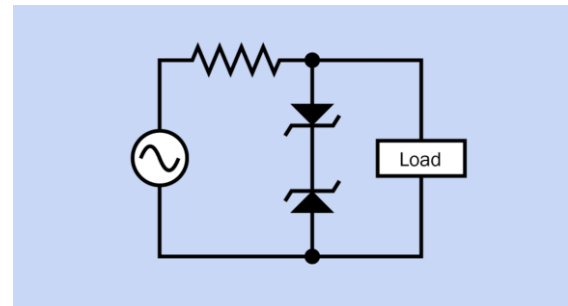


Figure 26-21. Two Zener diodes placed in series, with opposite polarities, can clip or limit the voltage sinewave of an AC signal.

Voltage Sensing

A Zener diode can be used to sense a small shift in voltage and provide a switched output in response.

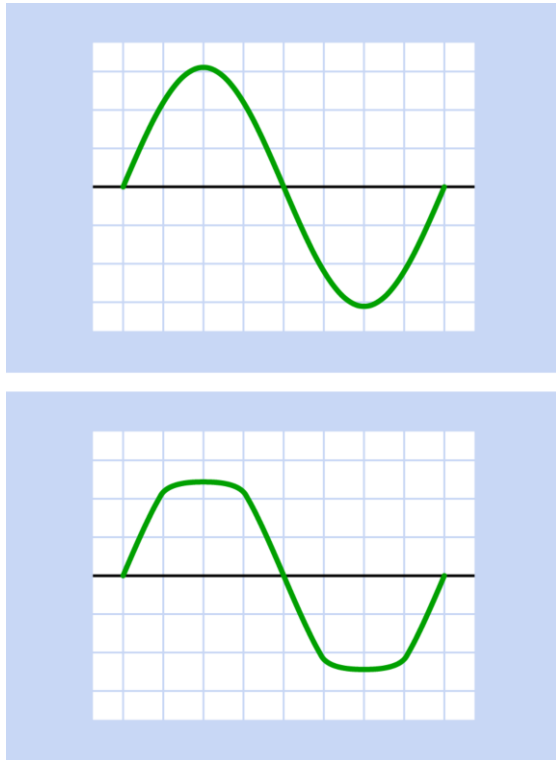


Figure 26-22. AC input showing a pure sinewave (left) and a clipped version (right) created by Zener diodes wired in series, as in the previous figure.

In [Figure 26-23](#), the upper schematic shows a Zener diode preventing voltage from reaching the emitter of a PNP transistor while the divided input signal is below the Zener (breakdown) voltage of the diode. In this mode, the transistor is relatively non-conductive, very little current flows through it, and the output is now at near-zero voltage. As soon as the input signal rises above the Zener voltage, the transistor switches on and power is supplied to the output. The input is thus replicated in the output, as shown in the upper portion of [Figure 26-24](#).

In [Figure 26-23](#), the lower schematic shows a Zener diode preventing voltage from reaching the base of an NPN transistor while the input signal is below the Zener (breakdown) voltage of the diode. In this mode, the transistor is relatively non-conductive, and power is supplied to the output. As soon as the input signal rises above

the Zener voltage, the transistor is activated, diverting the current to ground and bypassing the output, which is now at near-zero voltage. The input is thus inverted, as shown in the lower portion of [Figure 26-24](#) (provided there is enough current to drive the transistor into saturation).

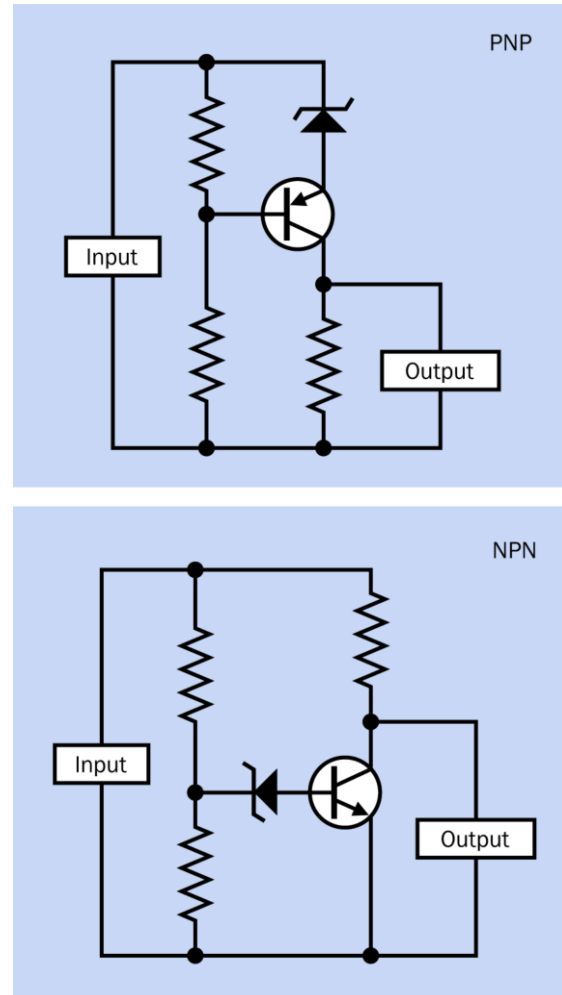


Figure 26-23. A Zener diode can be used in conjunction with a PNP transistor. See text for details.

What Can Go Wrong

Overload

If maximum forward current is exceeded, the heat generated is likely to destroy the diode. If the diode is reverse-biased beyond its peak in-

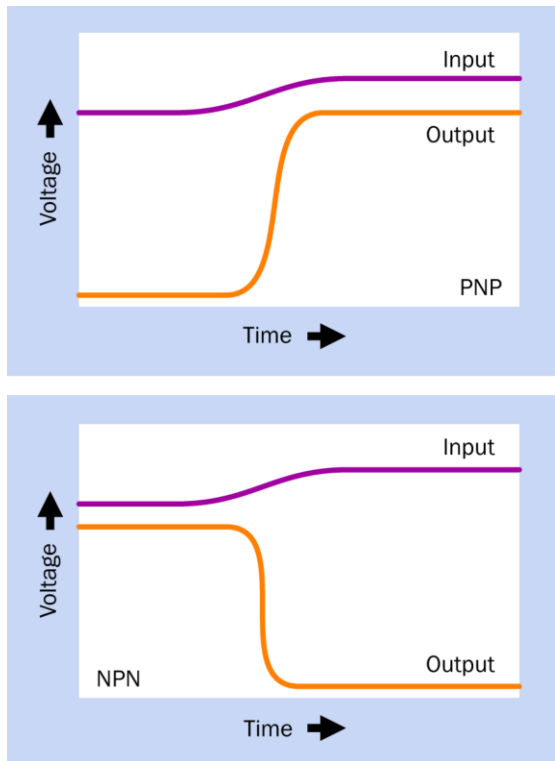


Figure 26-24. Theoretical output from the transistors in the two previous schematics.

verse voltage limit, the current will overwhelm the diode's ability to block it, and an *avalanche breakdown* will occur, once again probably destroying the component. The graph in [Figure 26-5](#) illustrates the performance range of a hypothetical generic diode.

Reversed Polarity

Zener diodes look almost identical to other types, and all diodes share the same convention of marking the cathode for identification. Yet Zeners must be reverse-biased while others are forward-biased. This creates a significant risk of installing a diode “the wrong way around,” with potentially destructive or at least confusing results, especially when used in a power supply. The very low resistance of a diode to forward current makes it especially vulnerable to burnout if installed incorrectly.

Wrong Type of Diode

If a Zener diode is used accidentally where a signal or rectifier diode is appropriate, the circuit will malfunction, as the Zener will probably have a much lower breakdown voltage, and therefore will not block reverse current. Conversely, if a signal or rectifier diode is used where the circuit calls for a Zener diode, reverse voltage will be clamped (or regulated at the diode's forward voltage value). Since diodes are often poorly marked, a sensible precaution is to store Zener diodes separately from all other types.

unijunction transistor

The unijunction transistor (UJT) and programmable unijunction transistor (PUT) are different internally, but are sufficiently similar in function to be combined in this entry.

OTHER RELATED COMPONENTS

- **diode** (See [Chapter 26](#))
- **bipolar transistor** (See [Chapter 28](#))
- **field-effect transistor** (See [Chapter 29](#))

What It Does

Despite their names, the unijunction transistor (UJT) and programmable unijunction transistor (PUT) are not current-amplification devices like bipolar transistors. They are switching components that are more similar to diodes than to transistors.

The UJT can be used to build low- to mid-frequency oscillator circuits, while the PUT provides similar capability with the addition of more sophisticated control, and is capable of functioning at lower currents. The UJT declined in popularity during the 1980s after introduction of components such as the [555 timer](#), which offered more flexibility and a more stable output frequency, eventually at a competitive price. UJTs are now uncommon, but PUTs are still available in quantity as through-hole discrete components. Whereas an integrated circuit such as a 555 timer generates a square wave, unijunction transistors in oscillator circuits generate a series of voltage spikes.

The PUT is often used to trigger a **thyristor** (described in Volume 2) and has applications in low-power circuits, where it can draw as little as a few microamps.

Schematic symbols for the two components are shown in [Figure 27-1](#) and [Figure 27-2](#). Although the symbol for the UJT is very similar to the symbol for a field-effect transistor (FET), its behavior is quite different. The bent arrow identifies the UJT, while a straight arrow identifies the FET. This difference is of significant importance.

The schematic symbol for a PUT indicates its function, as it resembles a diode with the addition of a gate connection.

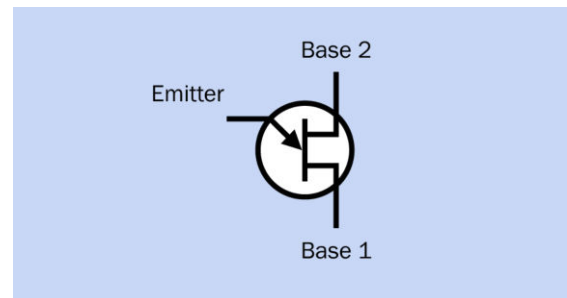


Figure 27-1. Schematic symbol for a unijunction transistor (UJT). Note the bent arrow. The symbol for a field-effect transistor looks similar, but has a straight arrow. The functionality of the two components is very different.

In [Figure 27-3](#), the transistors at left and center are old-original unijunction transistors, while the one at right is a programmable unijunction transistor. (Left: Maximum 300mW, 35V interbase voltage. Center: 450mW, 35V interbase voltage. Right: 300mW, 40V gate-cathode forward voltage, 40V anode-cathode voltage.)

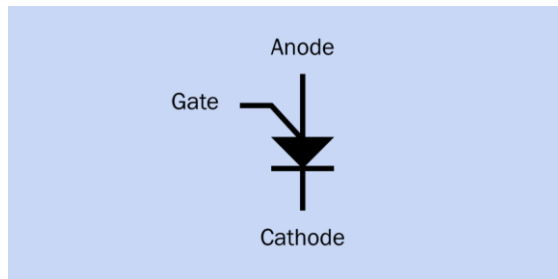


Figure 27-2. Schematic symbol for a programmable unijunction transistor (PUT). The symbol accurately suggests the similarity in function to a diode, with the addition of a gate to adjust the threshold voltage.

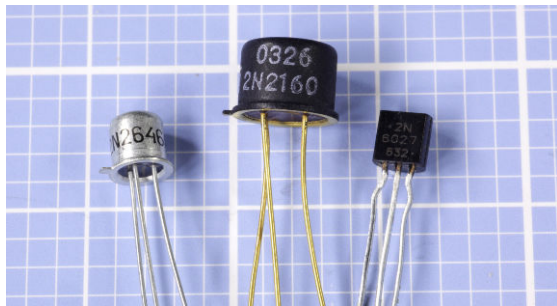


Figure 27-3. The unijunction transistors at left and center are becoming obsolete; the one at the right is a programmable unijunction transistor (PUT), still readily available and widely used as a thyristor trigger.

How It Works

The UJT is a three-terminal semiconductor device, but contains only two sections sharing a single junction—hence its name. Leads attached to opposite ends of a single channel of N-type semiconductor are referred to as base 1 and base 2, with base 2 requiring a slightly higher potential than base 1. A smaller P-type insert, midway between base 1 and base 2, is known as the emitter.

The diagram in [Figure 27-4](#) gives an approximate idea of internal function.

When no voltage is applied to the emitter, a relatively high resistance (usually more than 5K) prevents significant current flow from base 2 to base 1. When the positive potential at the emitter increases to a *triggering voltage* (similar to the *junction threshold voltage* of a forward-biased diode), the internal resistance of the UJT drops very rapidly, allowing current to enter the component via both the emitter and base 2, exiting at base 1. (The term “current” refers, here, to conventional current; electron flow is opposite.) Current flowing from base 2 to base 1 is significantly greater than current flowing from the emitter to base 1.

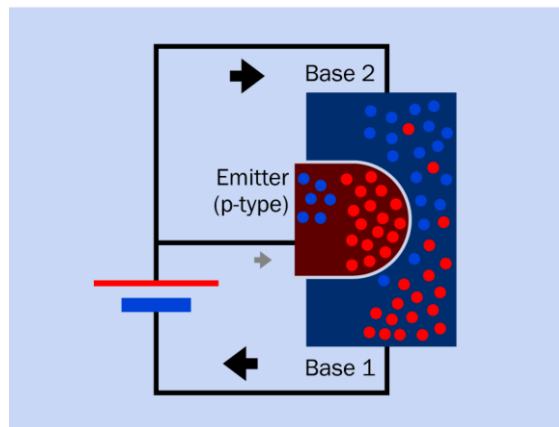
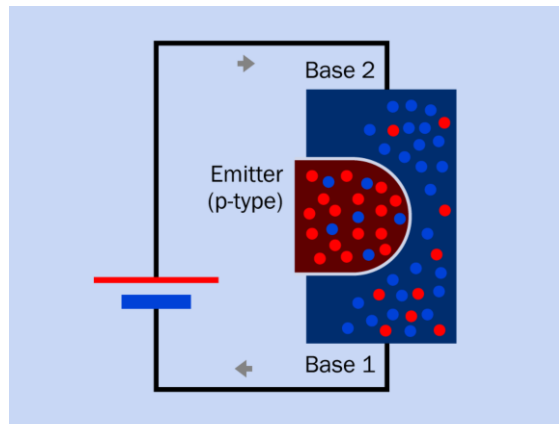


Figure 27-4. Internal workings of a unijunction transistor.

The graph in Figure 27-5 outlines the behavior of a UJT. As the voltage applied to the emitter increases, current flowing into the component from the emitter increases slightly, until the triggering voltage is reached. The component's internal resistance now drops rapidly. This pulls down the voltage at the emitter, while the current continues to increase significantly. Because of the drop in resistance, this is referred to as a *negative resistance* region. The resistance actually cannot fall below zero, but its change is negative. After emitter voltage drops to a minimum known as the *valley voltage*, the current continues to increase with a small increase in voltage. On datasheets, the peak current is often referred to as I_p while valley current is I_v .

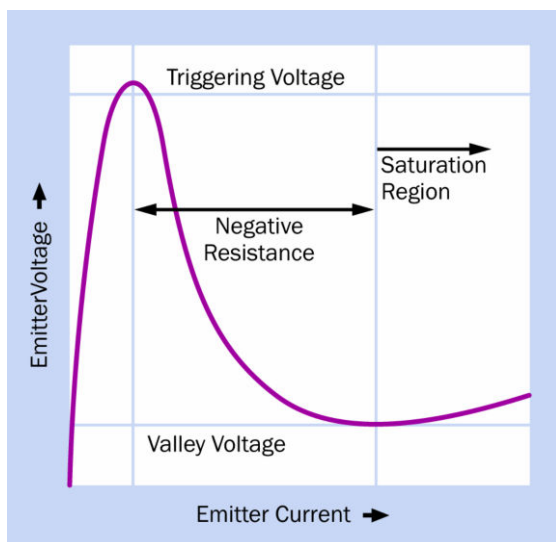


Figure 27-5. Response curve of a unijunction transistor (UJT). When positive potential at the emitter reaches the triggering voltage, internal resistance drops radically and the component goes through a phase known as “negative resistance” as current increases.

Figure 27-6 shows a test circuit to demonstrate the function of a UJT, with a volt meter indicating its status. A typical supply voltage would range from 9VDC to 20VDC.

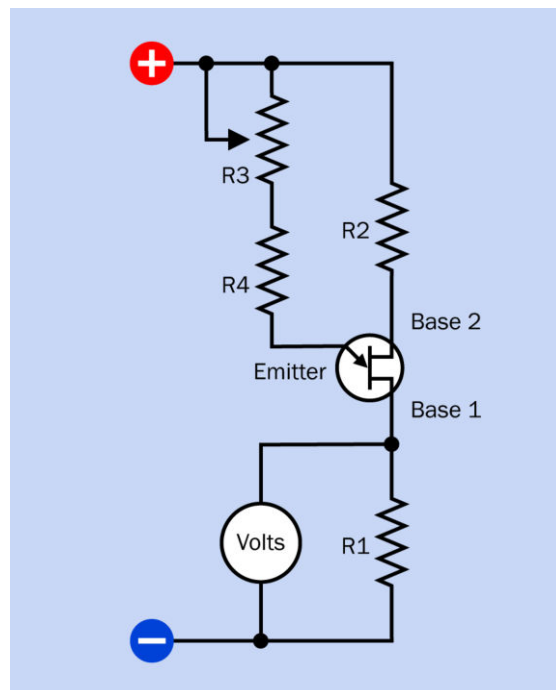


Figure 27-6. A test circuit for a unijunction transistor (UJT) using a volt meter to show its response as a potentiometer increases the voltage applied to its emitter.

A PUT behaves similarly in many ways to a UJT but is internally quite different, consisting of four semiconducting layers and functioning similarly to a thyristor.

The PUT is triggered by increasing the voltage on the anode. Figure 27-7 shows a test circuit for a PUT. This component is triggered when the voltage at its anode exceeds a threshold level, while the gate sets the threshold where this occurs. When the PUT is triggered, its internal resistance drops, and current flows from anode to cathode, with a smaller amount of current entering through the gate. This behavior is almost identical to that of a forward-biased diode, except that the threshold level can be controlled, or “programmed,” according to the value of the positive potential applied at the gate, with R1 and R2 establishing that potential by functioning as a voltage divider.

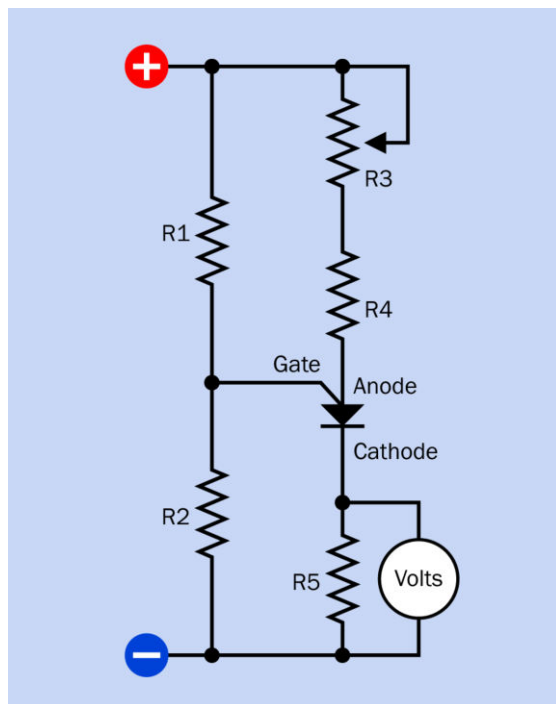


Figure 27-7. A test circuit for a programmable unijunction transistor (PUT) using a volt meter to show its response as a potentiometer increases the voltage applied to its anode.

The voltage output of a PUT follows a curve that is very similar to that shown in [Figure 27-5](#), although current and voltage would be measured at the cathode.

Variants

PUTs and UJTs are not made as surface-mount components.

UJTs are usually packaged in black plastic, although older variants were manufactured in cans. PUTs are almost all packaged in black plastic. With the leads pointing downward and the flat side facing toward the viewer, the lead functions of a PUT are usually anode, gate, and cathode, reading from left to right.

Values

The triggering voltage of a UJT can be calculated from the values of R1 and R2 in [Figure 27-7](#) and the voltage at base 1. The term R_{bb} is often used to represent the sum of $R1 + R2$, with V_{bb} representing the total voltage across the two resistors (this is the same as the supply voltage in [Figure 27-6](#)). V_t , the triggering voltage, is given by:

$$V_t = V_{bb} * (R1 / R_{bb})$$

The term $(R1/R_{bb})$ is known as the *standoff ratio*, often represented by the Greek letter ρ .

Typically the standoff ratio in a UJT is at least 0.7, as R1 is chosen to be larger than R2. Typical values for R1 and R2 could be 180 Ω and 100 Ω , respectively. If R4 is 50K and a 100K linear potentiometer is used for R3, the PUT should be triggered when the potentiometer is near the center of its range. The emitter saturation voltage is typically from 2V to 4V.

If using a PUT, typical values in the test circuit could be supply voltage ranging from 9VDC to 20VDC, with resistances 28K for R1 and 16K for R2, 20 Ω for R5, 280K for R4, and a 500K linear potentiometer for R3. The PUT should be triggered when the potentiometer is near the center of its range.

Sustained forward current from anode to cathode is usually a maximum of 150mA, while from gate to cathode the maximum is usually 50mA. Power dissipation should not exceed 300mW. These values should be lower at temperatures above 25 degrees Centigrade.

Depending on the PUT being used, power consumption can be radically decreased by upping the resistor values by a multiple of 100, while supply voltage can be decreased to 5V. The cathode output from the PUT would then be connected with the base of an NPN transistor for amplification.

How to Use it

Figure 27-8 shows a simple oscillator circuit built around a UJT, Figure 27-9 shows a comparable circuit for a PUT. Initially the supply voltage charges the capacitor, until the potential at the emitter of the UJT or the gate of the PUT reaches the threshold voltage, at which point the capacitor discharges through the emitter and the cycle repeats. Resistor values would be similar to those used in the test circuits previously described, while a capacitor value of $2.2\mu\text{F}$ would provide a visible pulse of the LED. Smaller capacitor values would enable faster oscillation. In the PUT circuit, adjusting the values of $R1$ and $R2$ would allow fine control of triggering the semiconductor.

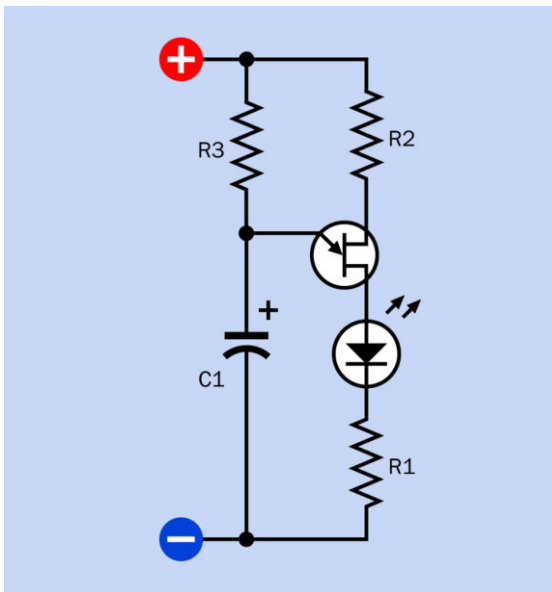


Figure 27-8. A basic oscillator circuit using a unijunction transistor (UJT). As the capacitor accumulates charge, the voltage on the emitter increases until it triggers the UJT, at which point the capacitor discharges through the emitter.

Probably the most common use for a PUT at this time is to trigger a **thyristor**.

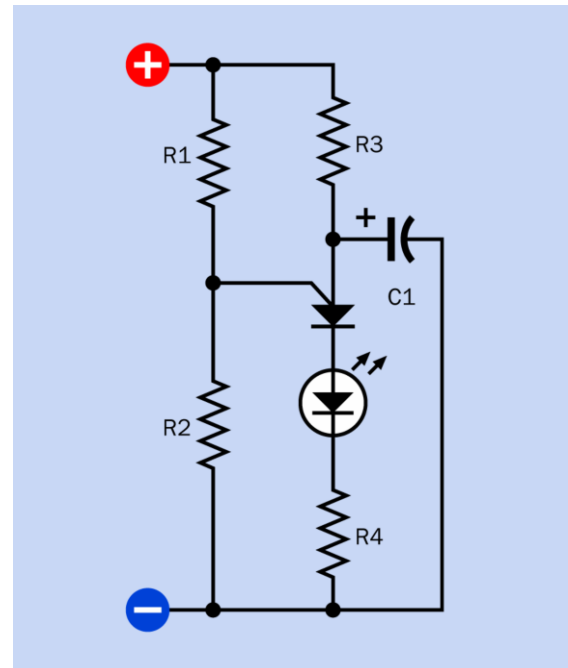


Figure 27-9. A basic oscillator circuit using a programmable unijunction transistor (PUT). As the capacitor accumulates charge, the voltage on the anode increases until it triggers the PUT, at which point the capacitor discharges through the anode. The gate voltage is preset by $R1$ and $R2$ to adjust the triggering voltage.

What Can Go Wrong

Name Confusion

A programmable unijunction transistor (PUT) is sometimes referred to simply as a “unijunction transistor” (UJT). Bearing in mind the totally different modes of operation of UJT and PUT, the PUT should always be identified by its acronym or by its full name. A circuit will not function if a UJT is substituted for a PUT, or a PUT is substituted for a UJT.

Incorrect Bias

Neither the UJT nor the PUT is designed to operate with reverse bias. In the UJT, a small forward bias should be applied from base 2 to base 1 (that is, base 2 should be at a higher potential relative to base 1) regardless of the voltage on the emitter. The emitter voltage may vary from 0 volts

upward. The PUT must be forward biased between its anode and cathode (the anode must have a higher potential relative to the cathode), with an intermediate positive voltage at the gate established by resistors R1 and R2 functioning as a voltage divider (see [Figure 27-7](#)). Failure to observe correct biasing will result in unpredictable behavior and possible damage to the component.

Overload

Like any semiconductor, the UJT and the PUT must be protected from excessive current, which can burn out the component. Never connect either of these components directly across a power source without appropriate resistances to limit current flow. Maximum continuous power dissipation for UJTs and PUTs is usually 300mW.

bipolar transistor

The word *transistor*, on its own, is often used to mean *bipolar transistor*, as this was the type that became most widely used in the field of discrete semiconductors. However, *bipolar transistor* is the correct term. It is sometimes referred to as a *bipolar junction transistor* or *BJT*.

OTHER RELATED COMPONENTS

- **unijunction transistor** (See [Chapter 27](#))
- **field-effect transistor** (See [Chapter 29](#))
- **diac** (Volume 2)
- **triac** (Volume 2)
- **relay** (Volume 2)
- **solid-state relay** (Volume 2)

What It Does

A bipolar transistor amplifies fluctuations in current or can be used to switch current on and off. In its amplifying mode, it replaced the [vacuum tubes](#) that were formerly used in the amplification of audio signals and many other applications. In its switching mode it resembles a **relay**, although in its “off” state the transistor still allows a very small amount of current flow, known as [leakage](#).

A bipolar transistor is described as a [discrete semiconductor device](#) when it is individually packaged, with three leads or contacts. A package containing multiple transistors is an [integrated circuit](#). A Darlington pair actually contains two transistors, but is included here as a discrete component because it is packaged similarly and functions like a single transistor. Most integrated circuits will be found in Volume 2 of this encyclopedia.

How It Works

Although the earliest transistors were fabricated from germanium, silicon has become the most commonly used material. Silicon behaves like an insulator, in its pure state at room temperature, but can be “doped” (carefully contaminated) with impurities that introduce a surplus of electrons unbonded from individual atoms. The result is an [N-type semiconductor](#) that can be induced to allow the movement of electrons through it, if it is [biased](#) with an external voltage. [Forward bias](#) means the application of a positive voltage, while [reverse bias](#) means reversing that voltage.

Other dopants can create a deficit of electrons, which can be thought of as a surplus of “holes” that can be filled by electrons. The result is a [P-type semiconductor](#).

A bipolar NPN transistor consists of a thin central P-type layer sandwiched between two thicker N-type layers. The three layers are referred to as [collector](#), [base](#), and [emitter](#), with a wire or contact

attached to each of them. When a negative charge is applied to the emitter, electrons are forced by mutual repulsion toward the central base layer. If a forward bias (positive potential) is applied to the base, electrons will tend to be attracted out through the base. However, because the base layer is so thin, the electrons are now close to the collector. If the base voltage increases, the additional energy encourages the electrons to jump into the collector, from which they will make their way to the positive current source, which can be thought of as having an even greater deficit of electrons.

Thus, the emitter of an NPN bipolar transistor emits electrons into the transistor, while the collector collects them from the base and moves them out of the transistor. It is important to remember that since electrons carry a negative charge, the flow of electrons moves from negative to positive. The concept of positive-to-negative current is a fiction that exists only for historical reasons. Nevertheless, the arrow in a transistor schematic symbol points in the direction of conventional (positive-to-negative) current.

In a PNP transistor, a thin N-type layer is sandwiched between two thicker P-type layers, the base is negatively biased relative to the emitter, and the function of an NPN transistor is reversed, as the terms “emitter” and “collector” now refer to the movement of electron-holes rather than electrons. The collector is negative relative to the base, and the resulting positive-to-negative current flow moves from emitter to base to collector. The arrow in the schematic symbol for a PNP transistor still indicates the direction of positive current flow.

Symbols for NPN and PNP transistors are shown in [Figure 28-1](#). The most common symbol for an NPN transistor is shown at top-left, with letters C, B, and E identifying collector, base, and emitter. In some schematics the circle in the symbols is omitted, as at top-right.

A PNP transistor is shown in the center. This is the most common orientation of the symbol, since its collector must be at a lower potential than its emitter, and ground (negative) is usually at the bottom of a schematic. At bottom, the PNP symbol is inverted, allowing the positions of emitter and collector to remain the same as in the symbol for the NPN transistor at the top. Other orientations of transistor symbols are often found, merely to facilitate simpler schematics with fewer conductor crossovers. The direction of the arrow in the symbol (pointing out or pointing in) always differentiates NPN from PNP transistors, respectively, and indicates current flowing from positive to negative.

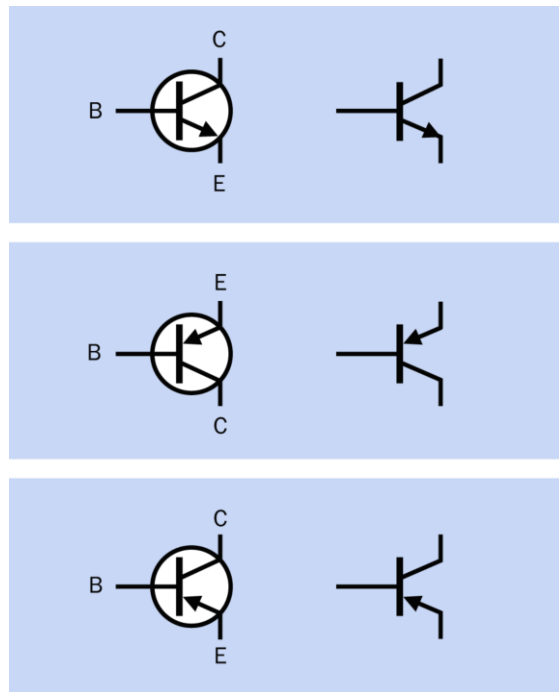


Figure 28-1. Symbols for an NPN transistor (top) and a PNP transistor (center and bottom). Depending on the schematic in which the symbol appears, it may be rotated or inverted. The circle may be omitted, but the function of the component remains the same.

NPN transistors are much more commonly used than PNP transistors. The PNP type was more difficult and expensive to manufacture initially, and

circuit design evolved around the NPN type. In addition, NPN transistors enable faster switching, because electrons have greater mobility than electron-holes.

To remember the functions of the collector and the emitter in an NPN transistor, you may prefer to think in terms of the collector collecting positive current *into* the transistor, and the emitter emitting positive current *out of* the transistor. To remember that the emitter is always the terminal with an arrow attached to it (both in NPN and PNP schematic symbols), consider that “emitter” and “arrow” both begin with a vowel, while “base” and “collector” begin with consonants. To remember that an NPN transistor symbol has its arrow pointing outward, you can use the mnemonic “N/ever P/ointing i/N.”

Current flow for NPN and PNP transistors is illustrated in Figure 28-2. At top-left, an NPN transistor passes no current (other than a small amount of leakage) from its collector to its emitter so long as its base is held at, or near, the potential of its emitter, which in this case is tied to negative or ground. At bottom-left, the purple positive symbol indicates that the base is now being held at a *relatively* positive voltage, at least 0.6 volts higher than the emitter (for a silicon-based transistor). This enables electrons to move from the emitter to the collector, in the direction of the blue arrows, while the red arrows indicate the conventional concept of current flowing from positive to negative. The smaller arrows indicate a smaller flow of current. A resistor is included to protect the transistor from excessive current, and can be thought of as the load in these circuits.

At top-right, a PNP transistor passes no current (other than a small amount of leakage) from its emitter to its collector so long as its base is held at, or near, the potential of the emitter, which in this case is tied to the positive power supply. At bottom-right, the purple negative symbol indicates that the base is now being held at a *relatively* negative voltage, at least 0.6 volts lower than the emitter. This enables electrons and current to flow as shown. Note that current flows

into the base in the NPN transistor, but out from the base in the PNP transistor, to enable conductivity. In both diagrams, the resistor that would normally be included to protect the base has been omitted for the sake of simplicity.

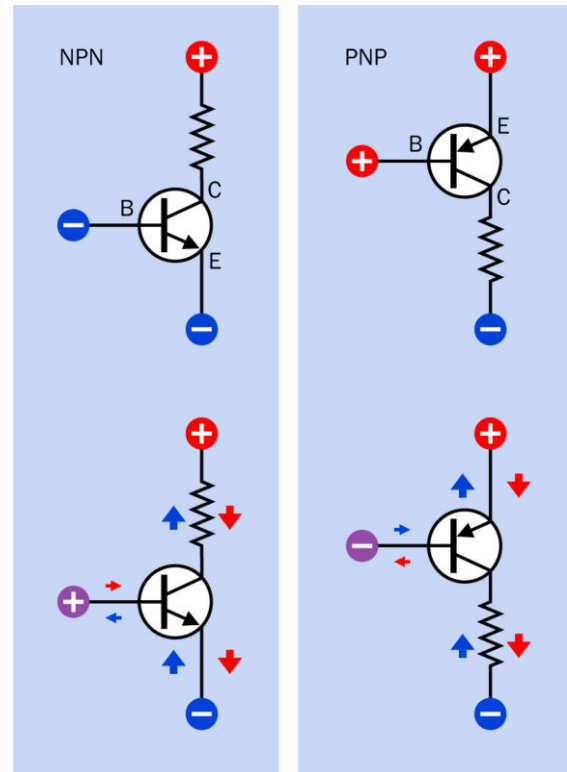


Figure 28-2. Current flow through NPN and PNP transistors. See text for details.

An NPN transistor amplifies its base current only so long as the positive potential applied to the collector is greater than the potential applied to the base, and the potential at the base must be greater than the potential at the emitter by at least 0.6 volts. So long as the transistor is biased in this way, and so long as the current values remain within the manufacturer’s specified limits, a small fluctuation in current applied to the base will induce a much larger fluctuation in current between the collector and the emitter. This is why a transistor may be described as a *current amplifier*.

A **voltage divider** is often used to control the base potential and ensure that it remains less than the potential on the collector and greater than the potential at the emitter (in an NPN transistor). See [Figure 28-3](#).

See [Chapter 10](#) for additional information about the function of a voltage divider.

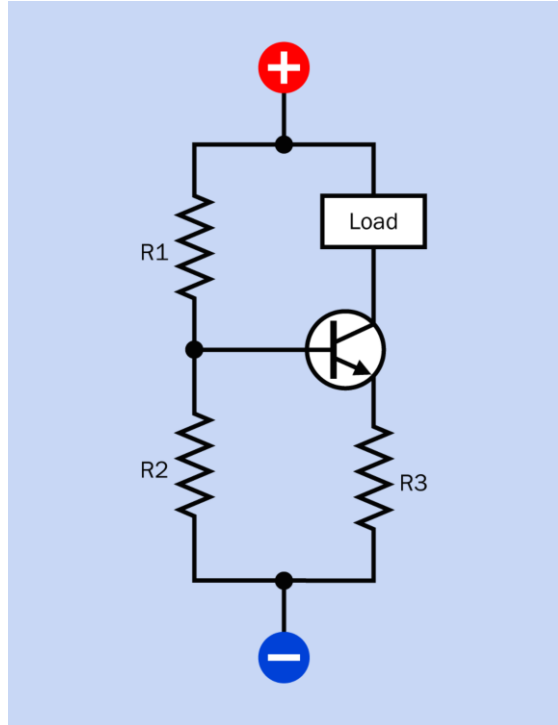


Figure 28-3. Resistors R1 and R2 establish a voltage divider to apply acceptable bias to the base of an NPN transistor.

Current Gain

The amplification of current by a transistor is known as its **current gain** or **beta value**, which can be expressed as the ratio of an increase in collector current divided by the increase in base current that enables it. Greek letter β is customarily used to represent this ratio. The formula looks like this:

$$\beta = \Delta I_c / \Delta I_b$$

where I_c is the collector current and I_b is the base current, and the Δ symbol represents a small change in the value of the variable that follows it.

Current gain is also represented by the term h_{FE} , where E is for the common Emitter, F is for Forward current, and lowercase letter h refers to the transistor as a “hybrid” device.

The beta value will always be greater than 1 and is often around 100, although it will vary from one type of transistor to another. It will also be affected by temperature, voltage applied to the transistor, collector current, and manufacturing inaccuracies. When the transistor is used outside of its design parameters, the formula to determine the beta value no longer directly applies.

There are only two connections at which current can enter an NPN transistor and one connection where it can leave. Therefore, if I_e is the current from the emitter, I_c is the current entering the collector, and I_b is the current entering the base:

$$I_e = I_c + I_b$$

If the potential applied to the base of an NPN transistor diminishes to the point where it is less than 0.6V above the potential at the emitter, the transistor will not conduct, and is in an “off” state, although a very small amount of leakage from collector to emitter will still occur.

When the current flowing into the base of the transistor rises to the point where the transistor cannot amplify the current any further, it becomes **saturated**, at which point its internal impedance has fallen to a minimal value. Theoretically this will allow a large flow of current; in practice, the transistor should be protected by resistors from being damaged by high current resulting from saturation.

Any transistor has maximum values for the collector current, base current, and the potential

difference between collector and emitter. These values should be provided in a datasheet. Exceeding them is likely to damage the component.

Terminology

In its *saturated mode*, a transistor's base is saturated with electrons (with no room for more) and the internal impedance between collector and emitter drops as low as it can go.

The *cutoff mode* of an NPN transistor is the state where a low base voltage eliminates all current flow from collector to emitter other than a small amount of leakage.

The *active mode*, or *linear mode*, is the intermediate condition between cutoff and saturated, where the beta value or h_{FE} (ratio of collector current to base current) remains approximately constant. That is, the collector current is almost linearly proportional to the base current. This linear relationship breaks down when the transistor nears its saturation point.

Variants

Small signal transistors are defined as having a maximum collector current of 500 mA and maximum collector power dissipation of 1 watt. They can be used for audio amplification of low-level inputs and for switching of small currents. When determining whether a small-signal transistor can control an inductive load such as a motor or relay coil, bear in mind that the initial current surge will be greater than the rated current draw during sustained operation.

Small switching transistors have some overlap in specification with small signal transistors, but generally have a faster response time, lower beta value, and may be more limited in their tolerance for collector current. Check the manufacturer's datasheet for details.

High frequency transistors are primarily used in video amplifiers and oscillators, are physically small, and have a maximum frequency rating as high as 2,000 MHz.

Power transistors are defined as being capable of handling at least 1 watt, with upper limits that can be as high as 500 watts and 150 amps. They are physically larger than the other types, and may be used in the output stages of audio amplifiers, and in switching **power supplies** (see [Chapter 16](#)). Typically they have a much lower current gain than smaller transistors (20 or 30 as opposed to 100 or more).

Sample transistors are shown in [Figure 28-4](#). Top: A 2N3055 NPN power transistor. This type was originally introduced in the late 1960s, and versions are still being manufactured. It is often found in power supplies and in push-pull power amplifiers, and has a total power dissipation rating of 115W. Second row, far left: general purpose switching-amplification PNP power transistor rated for up to 50W power dissipation. Second row, far right: A high-frequency switching transistor for use in lighting ballast, converters, inverters, switching regulators, and motor control systems. It tolerates relatively high voltages (up to 700V collector-emitter peak) and is rated for up to 80W total power dissipation. Second row, center-left and center-right: Two variants of the 2N2222 NPN small signal switching transistor, first introduced in the 1960s, and still very widely used. The metal can is the TO-18 package, capable of slightly higher power dissipation than the cheaper plastic TO-92 package (1.8W vs. 1.5W with a collector temperature no greater than 25 degrees Centigrade).

Packaging

Traditionally, small-signal transistors were packaged in small aluminum "cans" about 1/4" in diameter, and are still sometimes found in this form. More commonly they are embedded in buds of black plastic. Power transistors are packaged either in a rectangular module of black

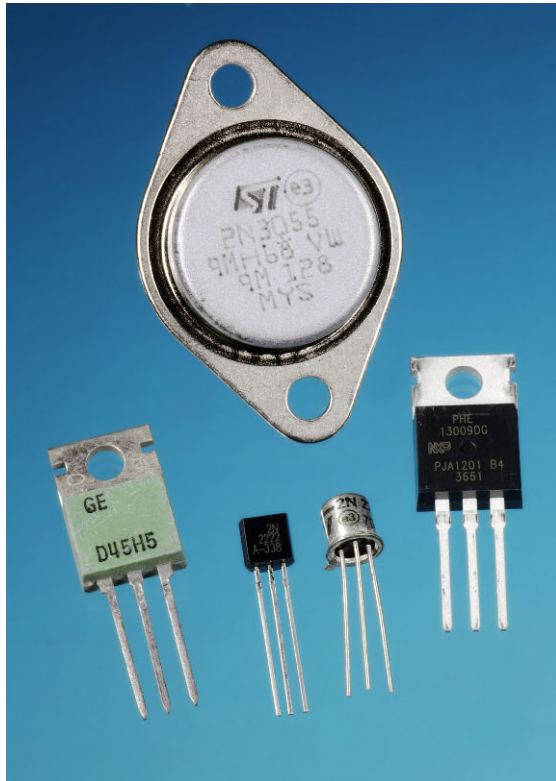


Figure 28-4. Samples of commonly used transistors. See text for details.

plastic with a metal back, or in a round metal “button.” Both of these forms are designed to dissipate heat by being screw-clamped to a *heat sink*.

Connections

Often a transistor package provides no clue as to which lead is the emitter, which lead is the base, and which lead is the collector. Old can-style packaging includes a protruding tab that usually points toward the emitter, but not always. Where power transistors are packaged in a metal enclosure, it is typically connected internally with the collector. In the case of surface-mount transistors, look for a dot or marker that should identify the base of a bipolar transistor or the gate of a field-effect transistor.

A through-hole transistor usually has its part number printed or engraved on its package, al-

though a magnifying glass may be necessary to see this. The component’s datasheet may then be checked online. If a datasheet is unavailable, meter-testing will be necessary to confirm the functions of the three transistor leads. Some multimeters include a transistor-test function, which may validate the functionality of a transistor while also displaying its beta value. Otherwise, a meter can be put in diode-testing mode, and an unpowered NPN transistor should behave as if diodes are connected between its leads as shown in [Figure 28-5](#). Where the identities of the transistor’s leads are unknown, this test will be sufficient to identify the base, after which the collector and emitter may be determined empirically by testing the transistor in a simple low-voltage circuit such as that shown in [Figure 28-6](#).

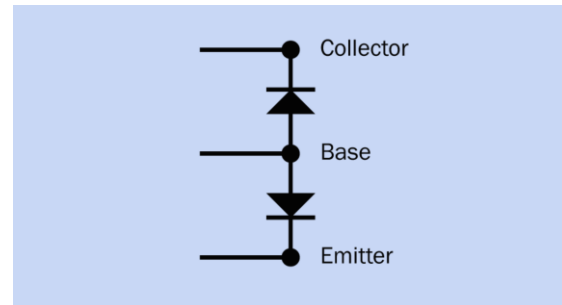


Figure 28-5. An NPN transistor can behave as if it contains two diodes connected as shown. Where the functions of the leads of the transistor are unknown, the base can be identified by testing for conductivity.

How to Use it

The following abbreviations and acronyms are common in transistor datasheets. Some or all of the letters following the initial letter are usually, but not always, formatted as subscripts:

h_{FE}

Forward current gain

β

Same as h_{FE}

V_{CEO}

Voltage between collector and emitter (no connection at base)

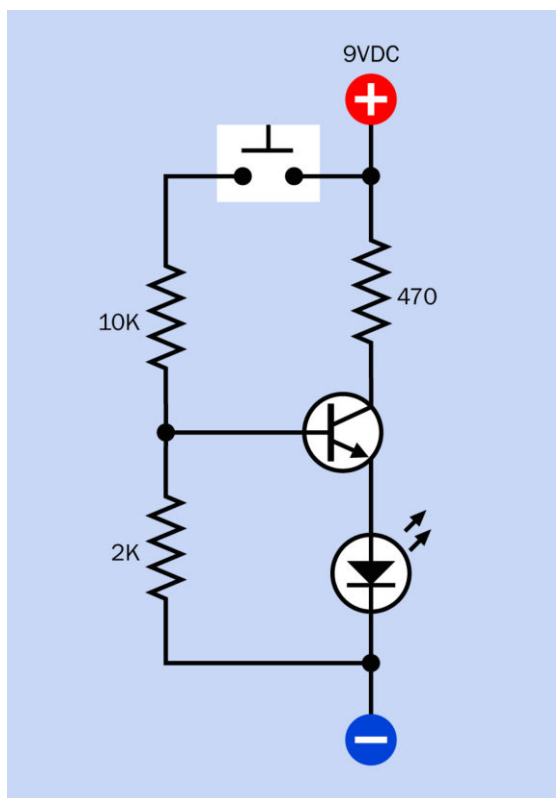


Figure 28-6. This simple schematic can be used to breadboard-test a transistor empirically, determining its functionality and the identities of its collector and emitter leads.

V_{CBO}

Voltage between collector and base (no connection at emitter)

V_{EBO}

Voltage between emitter and base (no connection at collector)

V_{CEsat}

Saturation voltage between collector and emitter

V_{BEsat}

Saturation voltage between base and emitter

I_c

Current measured at collector

I_{CM}

Maximum peak current at collector

I_{BM}

Maximum peak current at base

P_{TOT}

Total maximum power dissipation at room temperature

T_J

Maximum junction temperature to avoid damage

Often these terms are used to define “absolute maximum values” for a component. If these maximums are exceeded, damage may occur.

A manufacturer’s datasheet may include a graph showing the *safe operating area* (SOA) for a transistor. This is more common where power transistors are involved, as heat becomes more of an issue. The graph in Figure 28-7 has been adapted from a datasheet for a silicon diffused power transistor manufactured by Philips. The safe operating area is bounded at the top by a horizontal segment representing the maximum safe current, and at the right by a vertical segment representing the maximum safe voltage. However, the rectangular area enclosed by these lines is reduced by two diagonal segments representing the *total power limit* and the *second breakdown limit*. The latter refers to the tendency of a transistor to develop internal localized “hot spots” that tend to conduct more current, which makes them hotter, and able to conduct better—ultimately melting the silicon and causing a short circuit. The total power limit and the second breakdown limit reduce the safe operating area, which would otherwise be defined purely by maximum safe current and maximum safe voltage.

Uses for discrete transistors began to diminish when integrated circuits became cheaper and started to subsume multi-transistor circuits. For instance, an entire 5-watt audio amplifier, which used to be constructed from multiple components can now be bought on a chip, requiring just

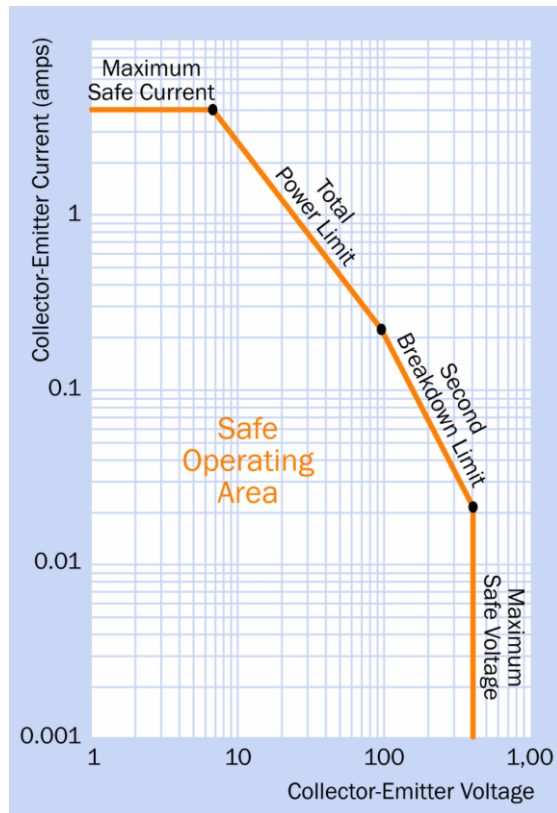


Figure 28-7. Adapted from a Philips datasheet for a power transistor, this graph defines a *safe operating area* (SOA) for the component. See text for details.

a few external capacitors. More powerful audio equipment typically uses integrated circuits to process inputs, but will use individual power transistors to handle high-wattage output.

Darlington Pairs

Discrete transistors are useful in situations where current amplification or switching is required at just one location in a circuit. An example would be where one output pin from a **microcontroller** must switch a small motor on and off. The motor may run on the same voltage as the microcontroller, but requires considerably more current than the typical 20mA maximum available from a microcontroller output. A *Darlington pair* of transistors may be used in this application. The overall gain of the pair can be 100,000 or more. See [Figure 28-8](#). If a power source feeding

through a potentiometer is substituted for the microcontroller chip, the circuit can function as a motor speed control (assuming that a generic DC motor is being used).

In the application shown here, the microcontroller chip must share a common ground (not shown) with the transistors. The optional resistor may be necessary to prevent leakage from the first transistor (when in its “off” state) from triggering the second. The diode protects the transistors from voltage transients that are likely when the motor stops and starts.

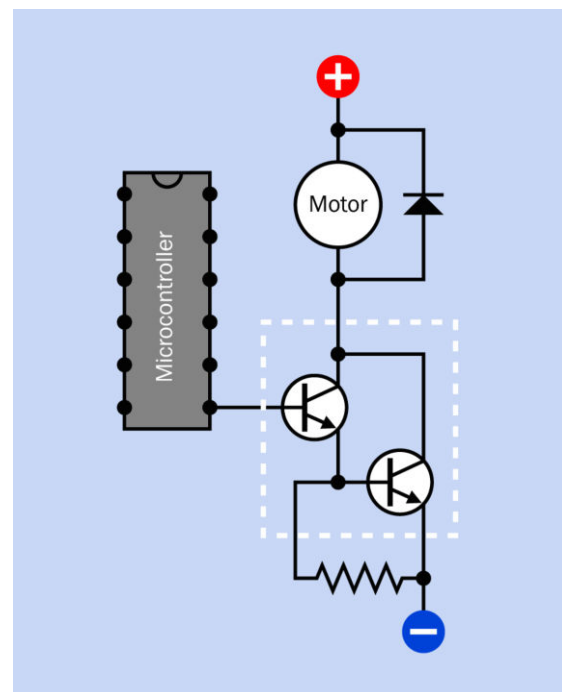


Figure 28-8. Where the emitter of one NPN transistor is coupled to the base of another, they form a *Darlington pair* (identified by the dashed rectangle in this schematic). Multiplying the gain of the first transistor by the gain of the second gives the total gain of the pair.

A Darlington pair can be obtained in a single transistor-like package, and may be represented by the schematic symbol shown in [Figure 28-9](#).

Various through-hole Darlington packages are shown in [Figure 28-10](#).

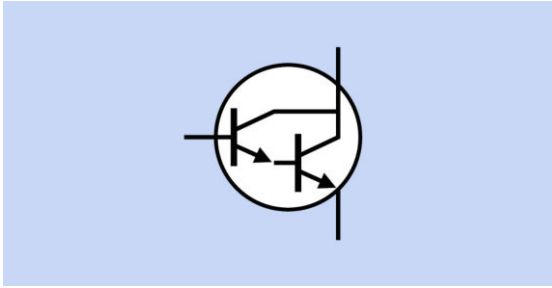


Figure 28-9. When a Darlington pair is embedded in a single transistor-like package, it may be represented by this schematic symbol. The leads attached to the package can be used as if they are the emitter, base, and collector of a single NPN transistor.

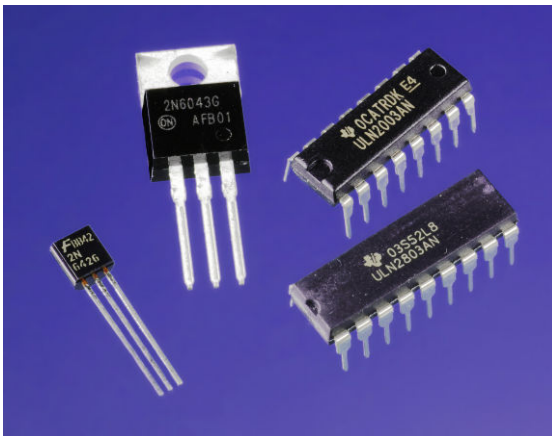


Figure 28-10. Various packaging options for Darlington pairs. From left to right: The 2N6426 contains a Darlington pair rated to pass up to 500mA continuous collector current. The 2N6043 is rated for 8A continuous. The ULN2003 and ULN2003AN chips contain seven and eight Darlington pairs, respectively.

Seven or eight Darlington pairs can be obtained in a single integrated chip. Each transistor pair in these chips is typically rated at 500mA, but they can be connected in parallel to allow higher currents. The chip usually includes protection diodes to allow it to drive inductive loads directly.

A typical schematic is shown in Figure 28-11. In this figure, the microcontroller connections are hypothetical and do not correspond with any actual chip. The Darlington chip is a ULN2003 or similar, containing seven transistor pairs, each with an “input” pin on the left and an “output”

pin opposite it on the right. Any of pins 1 through 7 down the left side of the chip can be used to control a device connected to a pin on the opposite side.

A high input can be thought of as creating a negative output, although in reality the transistors inside the chip are sinking current via an external device—a motor, in this example. The device can have its own positive supply voltage, shown here as 12VDC, but must share a common ground with the microcontroller, or with any other component which is being used on the input side. The lower-right pin of the chip shares the 12VDC supply because this pin is attached internally to clamp diodes (one for each Darlington pair), which protect against surges caused by inductive loads. For this reason, the motor does not have a clamp diode around it in the schematic.

The Darlington chip does not have a separate pin for connection with positive supply voltage, because the transistors inside it are sinking power from the devices attached to it.

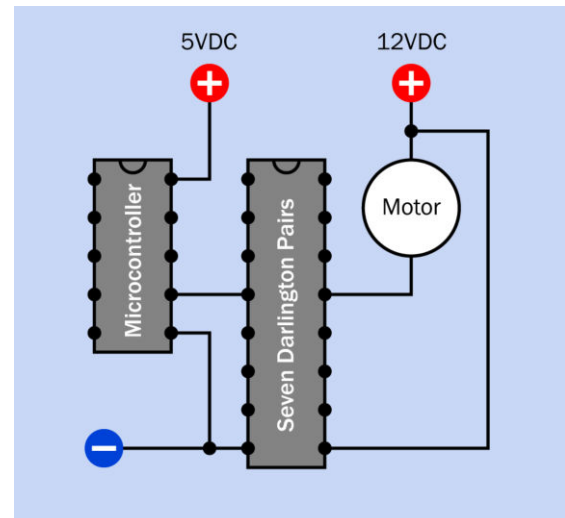


Figure 28-11. A chip such as the ULN2003 contains seven Darlington pairs. It will sink current from the device it is driving. See text for details.

A surface-mount Darlington pair is shown in Figure 28-12. This measures just slightly more

than 0.1" long but is still rated for up to 500mA collector current or 250mW total power dissipation (at a component temperature no higher than 25 degrees Centigrade).

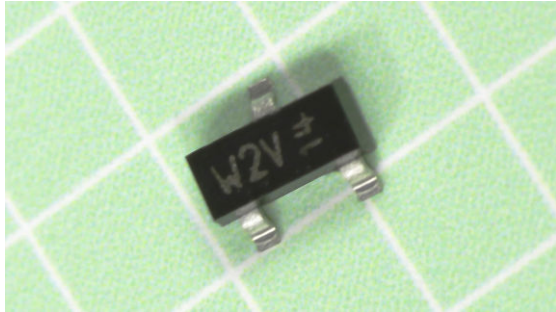


Figure 28-12. A surface-mount package for a Darlington pair. Each square in the background grid measures 0.1". See text for additional details.

Amplifiers

Two basic types of transistor amplifiers are shown in Figure 28-13 and Figure 28-14. The *common-collector* configuration has current gain but no voltage gain. The capacitor on the input side blocks DC current from entering the amplifier circuit, and the two resistors forming a voltage divider on the base of the transistor establish a voltage midpoint (known as the *quiescent point* or *operating point*) from which the signal to be amplified may deviate above and below.

The *common-emitter* amplifier provides voltage gain instead of current gain, but inverts the phase of the input signal. Additional discussion of amplifier design is outside the scope of this encyclopedia.

In switching applications, modern transistors have been developed to handle a lot of current compared with earlier versions, but still have some limitations. Few power transistors can handle more than 50A flowing from collector to emitter, and 1,000V is typically a maximum value. Electromechanical relays continue to exist because they retain some advantages, as shown in the table in Figure 28-15, which compares switching capabilities of transistors, **solid-state relays**, and electromechanical **relays**.

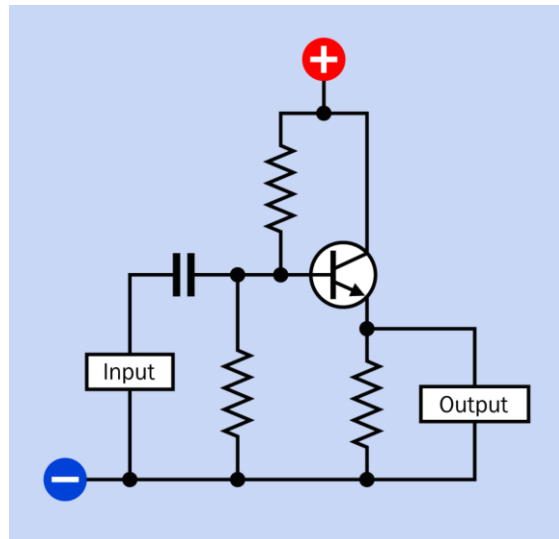


Figure 28-13. The basic schematic for a common-collector amplifier. See text for details.

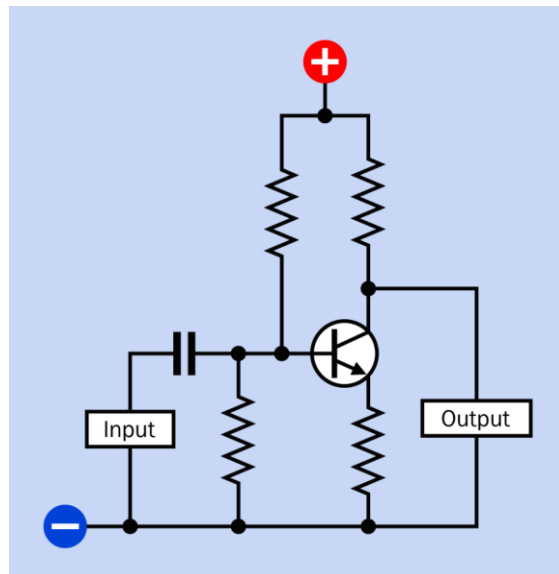


Figure 28-14. The basic schematic for a common-emitter amplifier. See text for details.

	Transistor	Solid-State Relay	Electro-Mechanical Relay
Long-term reliability	Excellent	Excellent	Limited
NC contacts?	No	Yes	Yes
DT contacts?	No	No	Yes
Ability to switch large current	Limited	Some	Good
Ability to switch AC	No	Yes	Yes
Can be triggered by AC voltage?	No	Yes	Yes
Suitability for miniaturization	Excellent	Poor	Poor
Vulnerable to heat	Yes	Yes	Not very
Vulnerable to corrosion	No	No	Yes
OK at high speed	Excellent	Good	Poor
Price advantage for low-voltage low-current	Yes	No	No
Price advantage for high-voltage high-current	No	No	Yes
Current leakage when "off"	Yes	Yes	No
Trigger circuit isolated from switched circuit	No	Yes	Yes

Figure 28-15. A comparison of characteristics of switching devices.

What Can Go Wrong

Wrong Connections on a Bipolar Transistor

Failing to identify a transistor’s leads or contacts correctly can obviously be a potential source of damage, but swapping the collector and emitter

accidentally will not necessarily destroy the transistor. Because of the inherent symmetry of the device, it will in fact function with collector and emitter connections reversed. Rohm, a large semiconductor manufacturer, has included this scenario in its general information pages and concludes that the primary indicator of transposed connections is that the β value, or h_{FE} , drops to about 1/10th of specification. If you are using a transistor that works but provides much less amplification than you expect, check that the emitter and collector leads are not transposed.

Wrong Connections on a Darlington Pair Chip

While a single-component package for a Darlington pair functions almost indistinguishably from a single transistor, multiple Darlington pairs in a DIP package may create confusion because the component behaves differently from most other chips, such as logic chips.

A frequent error is to ground the output device instead of applying positive power to it. See [Figure 28-11](#) and imagine an erroneous connection of negative power instead of the 12VDC positive power.

Additional confusion may be caused by reading a manufacturer’s datasheet for a Darlington pair chip such as the ULN2003. The datasheet depicts the internal function of the chip as if it contains logic inverters. While the chip can be imagined as behaving this way, in fact it contains bipolar transistors that amplify the current applied to the base of each pair. The datasheet also typically will not show the positive connection that should be made to the common-diode pin (usually at bottom-right), to provide protection from surges caused by inductive loads. This pin must be distinguished carefully from the common-ground pin (usually at bottom-left). The positive connection to the common-diode pin is optional; the common-ground connection is mandatory.

Soldering Damage

Like any semiconductor, transistors are vulnerable to heat and can be damaged while soldering, although this seldom happens if a low-wattage iron is used. A copper alligator clip can be applied as a heat sink to each lead before it is soldered.

Excessive Current or Voltage

During use, a transistor will be damaged if it is subjected to current or voltage outside of its rated range. Passing current through a transistor without any series resistance to protect it will almost certainly burn it out, and the same thing can happen if incorrect resistor values are used.

The maximum wattage that a transistor can dissipate will be shown in its datasheet. Suppose, for example, this figure is 200mW, and you are using a 12VDC supply. Ignoring the base current, the maximum collector current will be $200 / 12 =$ approximately 15mA. If the transistor's emitter is connected to ground, and the load applied to the transistor output has a high impedance, and if we ignore the transresistance, Ohm's Law suggests that a resistor that you place between the collector and the supply voltage should have a resistance of at least $12 / 0.015 = 800$ ohms.

When transistors are used in switching applications, it is customary for the base current to be 1/5th of the collector current. In the example discussed here, a 4.7K resistor might be appropriate. A meter should be used to verify actual current and voltage values.

Excessive Leakage

In a Darlington pair, or any other configuration where the output from one transistor is connected with the base of another, leakage from the first transistor while in its "off" state can be amplified by the second transistor. If this is unacceptable, a bypass resistor can be used to divert some of the leakage from the base of the second transistor to ground. Of course the resistor will also steal some of the base current when the first transistor is active, but the resistor value is typically chosen so that it takes no more than 10% of the active current. See [Figure 28-8](#) for an example of a bypass resistor added to a Darlington pair.

field effect transistor

The term **field-effect transistor** encompasses a family primarily consisting of the *junction field-effect transistor* (or *JFET*, which is the simplest generic type) and the *metal-oxide semiconductor field-effect transistor* (or *MOSFET*, also sometimes known as an *insulated-gate field-effect transistor*, or *IGFET*). Because the principles of operation overlap considerably, the entire -FET family is grouped in this entry.

OTHER RELATED COMPONENTS

- **bipolar transistor** (See [Chapter 28](#))
- **unijunction transistor** (See [Chapter 27](#))
- **diode** (See [Chapter 26](#))

What It Does

A field-effect transistor creates an electric field to control current flowing through a channel in a semiconductor. MOSFETs of microscopic size form the basis of complementary metal oxide semiconductor (CMOS) *integrated circuit chips*, while large discrete MOSFETs are capable of switching substantial currents, in lamp dimmers, audio amplifiers, and motor controllers. FETs have become indispensable in computer electronics.

A **bipolar transistor** is generally thought of as a *current amplifier* because the current passing through it is controlled by a smaller amount of current passing through the base. By contrast, all FETs are considered to be *voltage amplifiers*, as the control voltage establishes field intensity, which requires little or no current. The negligible leakage through the gate of an FET makes it ideal for use in low-power applications such as portable hand-held devices.

How It Works

This section is divided into two subsections, describing the most widely used FETs: JFETs and MOSFETs.

JFETs

A *junction field-effect transistor* (or *JFET*) is the simplest form of FET. Just as a bipolar transistor can be of NPN or PNP type, a JFET can have an *N-channel* or *P-channel*, depending whether the channel that transmits current through the device is negatively or positively *doped*. A detailed explanation of semiconductor doping will be found in the **bipolar transistor** entry.

Because negative charges have greater mobility, the N-channel type allows faster switching and is more commonly used than the P-channel type. A schematic symbol for it is shown in [Figure 29-1](#) alongside the schematic for an NPN transistor. These symbols suggest the similarity of the devices as amplifiers or switches, but it is important to remember that the FET is a primarily a voltage amplifier while the bipolar transistor is a current amplifier.

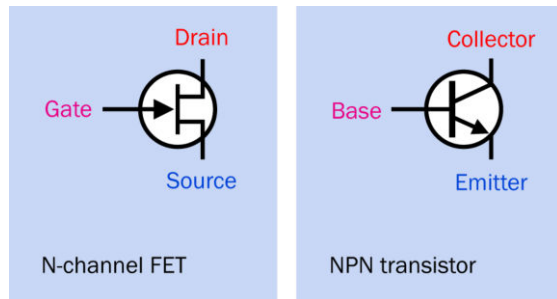


Figure 29-1. A comparison between schematic symbols for N-channel JFET (left) and NPN bipolar transistor (right) suggests their functional similarity as switches or amplifiers, although their behavior is markedly different.

Three JFETs are shown in Figure 29-2. The N-channel J112 type is supplied by several manufacturers, the figure showing two samples, one from Fairchild Semiconductor (left) and the other from On Semiconductor (right). Although the full part numbers are different, the specifications are almost identical, including a drain-gate voltage of 35V, a drain-source voltage of 35V, and a gate current of 50mA. The metal-clad 2N4392 in the center has similar values but is three times the price, with a much higher power dissipation of 1.8W, compared with 300mW and 350mW for the other two transistors respectively.

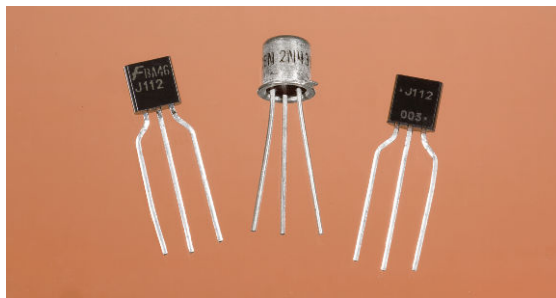


Figure 29-2. Junction Field Effect Transistors (JFETs). See text for details.

Schematic symbols for N-channel and P-channel JFETs are shown in Figure 29-3, N-channel being on the left while P-channel is on the right. The upper-left and lower-left symbol variants are both widely used and are functionally identical. The upper-right and lower-right variants likewise

mean the same thing. Because the upper variants are symmetrical, an S should be added to clarify which terminal is the source. In practice, the S is often omitted, allowing some ambiguity. While the source and drain of some JFETs are in fact interchangeable, this does not apply to all types.

The circle around each symbol is occasionally omitted when representing discrete components, and is almost always omitted when multiple FETs are shown connected to form an integrated circuit.

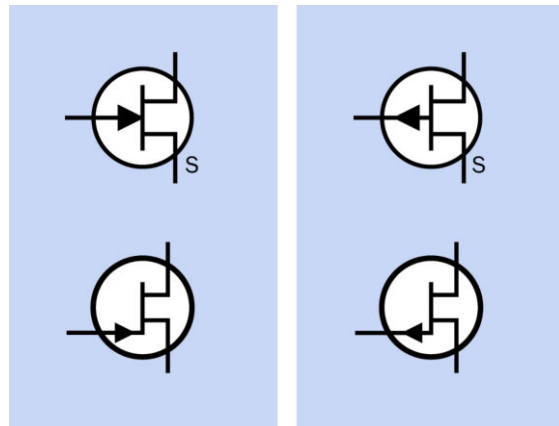


Figure 29-3. Schematic symbols for junction field-effect transistors (JFETs). Left: N-channel. Right: P-channel. The symbols at top and bottom on each side are functionally identical. Circles may be omitted. Letter S may be omitted from the symmetrical symbol variants, even though this creates some ambiguity.

The internal function of an N-channel JFET is shown diagrammatically in Figure 29-4. In this component, the source terminal is a source of electrons, flowing relatively freely through the N-doped channel and emerging through the drain. Thus, conventional current flows nonintuitively from the drain terminal, to the source terminal, which will be of lower potential.

The JFET is like a normally-closed switch. It has a low resistance so long as the gate is at the same potential as the source. However, if the potential of the gate is reduced below the potential of the source—that is, the gate acquires a more rela-

tively negative voltage than the source—the current flow is *pinched off* as a result of the field created by the gate. This is suggested by the lower diagram in Figure 29-4.

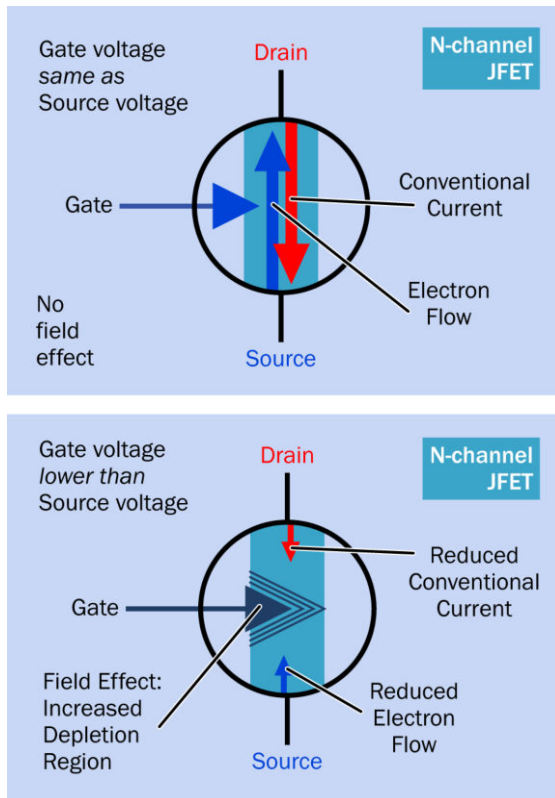


Figure 29-4. At top, conventional current flows freely from drain to source through the channel of an N-doped JFET. At bottom, the lowered voltage of the gate relative to the source creates a field effect that pinches off the flow of current.

The situation for a P-channel JFET is reversed, as shown in Figure 29-5. The source is now positive (but is still referred to as the source), while the drain can be grounded. Conventional current now flows freely from source to drain, so long as the gate is at the same positive potential as the source. If the gate voltage rises above the source voltage, the flow is pinched off.

A bipolar transistor tends to block current flow by default, but becomes less resistive when its base is forward-biased. Therefore it can be re-

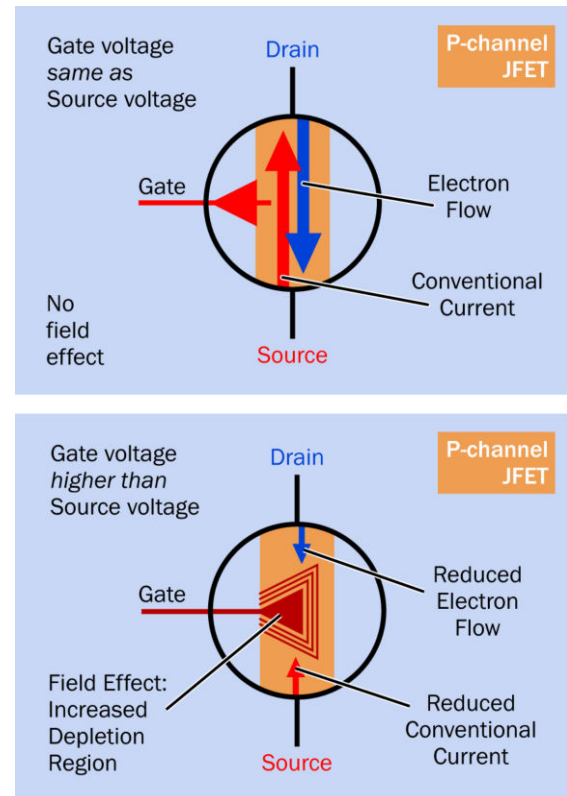


Figure 29-5. At top, conventional current flows freely from source to drain through the channel of a P-doped JFET. At bottom, the higher voltage of the gate relative to the source creates a field effect that pinches off the flow of current.

ferred to as an *enhancement device*. By contrast, an N-channel JFET allows current to flow by default, and becomes more resistive when its base is reverse-biased, which widens the *depletion layer* at the base junction. Consequently it can be referred to as a *depletion device*.

The primary characteristics of a junction field-effect transistor relative to an NPN bipolar transistor are summarized in the table in Figure 29-6.

JFET Behavior

The voltage difference between gate and source of a JFET is usually referred to as V_{gs} while the voltage difference between drain and source is referred to as V_{ds} .

	NPN Bipolar Transistor	N-Channel JFET
Type of amplifier	Current	Voltage
Active bias	Positive	Negative
Unbiased state	Nonconductive	Conductive
Biased state	More conductive	Less conductive

Figure 29-6. This table contrasts the characteristics of an N-channel JFET with those of an NPN bipolar transistor.

Suppose the gate of an N-channel JFET is connected with the source, so that $V_{gs}=0$. Now if V_{ds} increases, the current flowing through the channel of the JFET also increases, approximately linearly with V_{ds} . In other words, initially the JFET behaves like a low-value resistor in which the voltage across it, divided by the amperage flowing through it, is approximately constant. This phase of the JFET's behavior is known as its *ohmic region*. While the unbiased resistance of the channel in a JFET depends on the component type, it is generally somewhere between 10 Ω and 1K.

If V_{ds} increases still further, eventually no additional flow of current occurs. At this point the channel has become *saturated*, and this plateau zone is referred to as the *saturation region*, often abbreviated I_{dss} , meaning "the saturated drain current at zero bias." Although this is a nearly constant value for any particular JFET, it may vary somewhat from one sample of a component to another, as a result of manufacturing variations.

If V_{ds} continues to increase, the component finally enters a *breakdown* state, sometimes referred to by its full formal terminology as *drain-source breakdown*. The current passing through the JFET will now be limited only by capabilities

of the external power source. This breakdown state can be destructive to the component, and is comparable to the breakdown state of a typical **diode**.

What if the voltage at the gate is reduced below the voltage at the source—such as V_{gs} becomes negative? In its ohmic region, the component now behaves as if it has a higher resistance, and it will reach its saturation region at a lower current value (although around the same value for V_{ds}). Therefore, by reducing the voltage on the gate relative to the voltage at the source, the effective resistance of the component increases, and in fact it can behave as a *voltage-controlled resistor*.

The upper diagram in [Figure 29-7](#) shows this graphically. Below it, the corresponding graph for a P-channel JFET looks almost identical, except that the current flow is reversed and is pinched off as the gate voltage rises above the source voltage. Also, the breakdown region is reached more quickly with a P-channel JFET than with an N-channel JFET.

MOSFETs

MOSFETs have become one of the most widely used components in electronics, everywhere from computer memory to high-amperage switching power supplies. The name is an acronym for *metal-oxide semiconductor field-effect transistor*. A simplified cross-section of an N-channel MOSFET is shown in [Figure 29-8](#).

Two MOSFETs are shown in [Figure 29-9](#).

Like a JFET, a MOSFET has three terminals, identified as drain, gate, and source, and it functions by creating a field effect that controls current flowing through a channel. (Some MOSFETs have a fourth terminal, described later). However, it has a metal source and drain making contact with each end of the channel (hence the term "metal" in its acronym) and also has a thin layer of silicon dioxide (hence the term "oxide" in its acronym) separating the gate from the channel, thus raising the impedance at the gate to at least

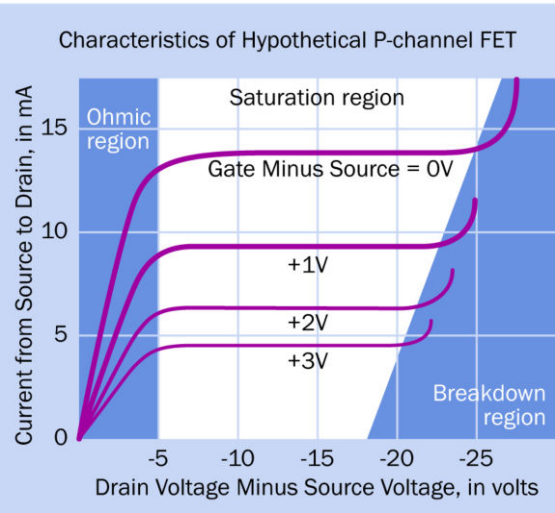
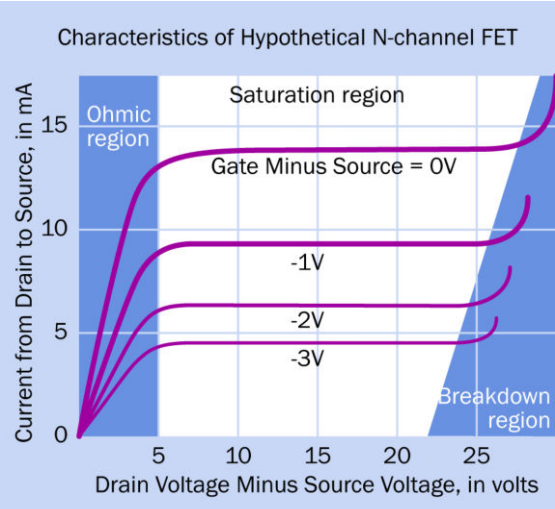


Figure 29-7. The top graph shows current passing through the channel of an N-channel JFET, depending on gate voltage and source voltage. The lower graph is for a P-channel JFET.

100,000 gigaohms and reducing gate current essentially to zero. The high gate impedance of a MOSFET allows it to be connected directly to the output of a digital integrated circuit. The layer of silicon dioxide is a *dielectric*, meaning that a field applied to one side creates an opposite field on the other side. The gate attached to the surface of the layer functions in the same way as one plate of a capacitor.

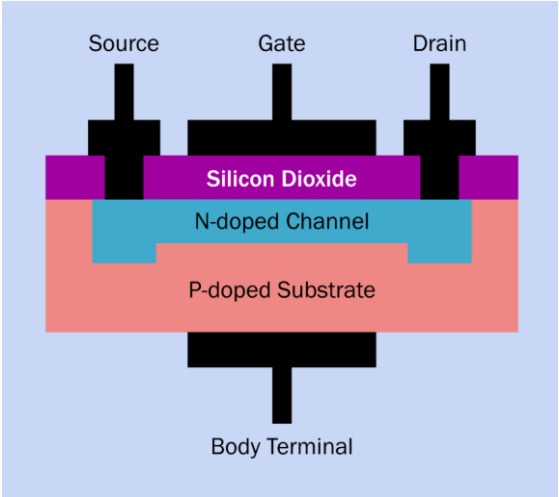


Figure 29-8. Simplified diagram of an N-channel MOSFET. The thickness of the silicon dioxide layer has been greatly exaggerated for clarity. The black terminals are metallic.



Figure 29-9. Two MOSFETs. At left, the TO-220 package claims a drain current of up to 65A continuous, and a drain-to-source breakdown voltage 100V. At right, the smaller package offers a drain current of 175mA continuous, and a drain-to-source breakdown voltage of 300V.

The silicon dioxide also has the highly desirable property of insulating the gate from the channel, thus preventing unwanted reverse current. In a JFET, which lacks a dielectric layer, if source voltage is allowed to rise more than about 0.6V higher than gate voltage, the direct internal connection between gate and channel allows negative

charges to flow freely from source to gate, and as the internal resistance will be very low, the resulting current can be destructive. This is why the JFET must always be reverse-biased.

A MOSFET is freed from these restrictions, and the gate voltage can be higher or lower than the source voltage. This property enables an N-channel MOSFET to be designed not only as a depletion device, but alternatively as an *enhancement device*, which is “normally off” and can be switched on by being forward-biased. The primary difference is the extent to which the channel in the MOSFET is N-doped with charge carriers, and therefore will or will not conduct without some help from the gate bias.

In a depletion device, the channel conducts, but applying negative voltage to the gate can pinch off the current.

In an enhancement device, the channel does not conduct, but applying positive voltage to the gate can make it start to do so.

In either case, a shift of bias from negative to positive encourages channel conduction; the depletion and enhancement versions simply start from different points.

This is clarified in [Figure 29-10](#). The vertical (logarithmic) scale suggests the current being conducted through the channel of the MOSFET, while the green curve describes the behavior of a depletion version of the device. Where this curve crosses the center line representing 0 volts bias, the channel is naturally conductive, like a JFET. Moving left down the curve, as reverse bias is applied (shown on the horizontal axis), the component becomes less conductive until finally its conductivity reaches zero.

Meanwhile on the same graph, the orange curve represents an enhancement MOSFET, which is nonconductive at 0 volts bias. As forward bias increases, the current also increases—similar to a bipolar transistor.

To make things more confusing, a MOSFET, like a JFET, can have a P-doped channel; and once

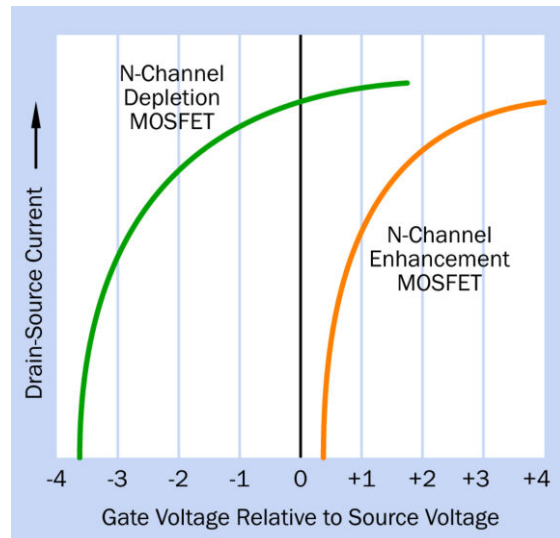


Figure 29-10. The current conduction of depletion and enhancement N-channel MOSFETs. See text for details. (Influenced by *The Art of Electronics* by Horowitz and Hill.)

again it can function in depletion or enhancement mode. The behavior of this variant is shown in [Figure 29-11](#). As before, the green curve shows the behavior of a depletion MOSFET, while the orange curve refers to the enhancement version. The horizontal axis now shows the voltage difference between the gate and the drain terminal. The depletion component is naturally conductive at zero bias, until the gate voltage increases above the drain voltage, pinching off the current flow. The enhancement component is not conductive until reverse bias is applied.

[Figure 29-12](#) shows schematic symbols that represent depletion MOSFETs. The two symbols on the left are functionally identical, representing N-channel versions, while the two symbols on the right represent P-channel versions. As in the case of JFETs, the letter “S” should be (but often is not) added to the symmetrical versions of the symbols, to clarify which is the source terminal. The left-pointing arrow identifies the components as N-channel, while in the symbols on the right, the right-pointing arrows indicate P-channel MOS-

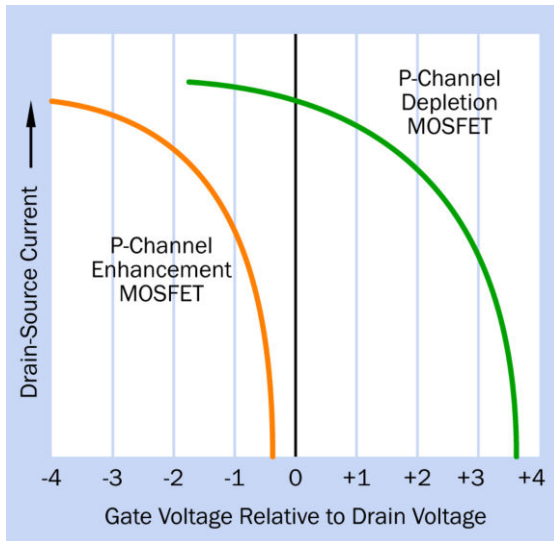


Figure 29-11. The current conduction of depletion and enhancement P-channel MOSFETs. See text for details.

FETs. The gap between the two vertical lines in each symbol suggests the silicon dioxide dielectric. The right-hand vertical line represents the channel.

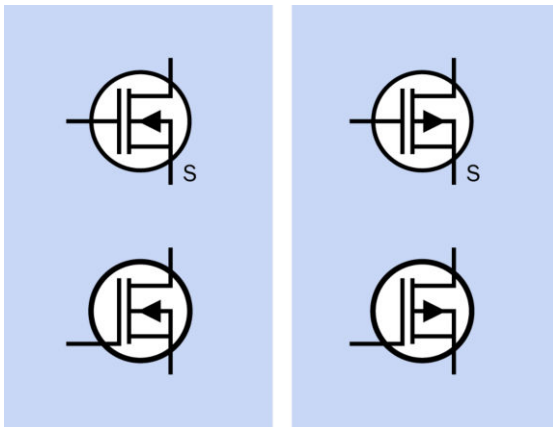


Figure 29-12. Schematic symbols for depletion MOSFETs. These function similarly to JFETs. The two symbols on the left are functionally identical, and represent N-channel depletion MOSFETs. The two symbols on the right are both widely used to represent P-channel depletion MOSFETs.

For enhancement MOSFETs, a slightly different symbol uses a broken line between the source and drain (as shown in Figure 29-13) to remind

us that these components are “normally off” when zero-biased, instead of “normally on.” Here again a left-pointing arrow represents an N-channel MOSFET, while a right-pointing arrow represents a P-channel MOSFET.

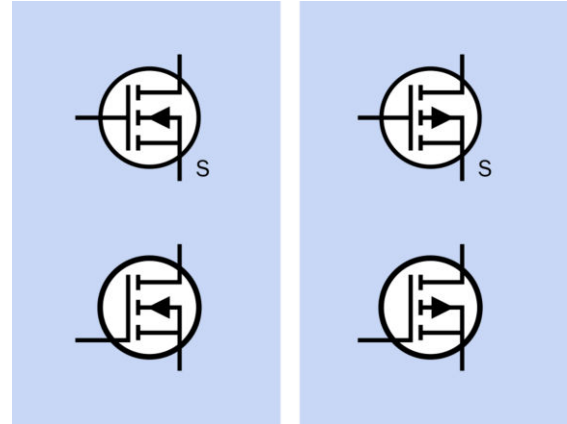


Figure 29-13. Schematic symbols for enhancement MOSFETs. The two on the left are functionally identical, and represent N-channel enhancement MOSFETs. The two on the right represent P-channel enhancement MOSFETs.

Because there is so much room for confusion regarding MOSFETs, a summary is presented in Figure 29-14 and Figure 29-15. In these figures, the relevant parts of each schematic symbol are shown disassembled alongside text explaining their meaning. Either of the symbols in Figure 29-14 can be superimposed on either of the symbols in Figure 29-15, to combine their functions. So, for instance, if the upper symbol in Figure 29-14 is superimposed on the lower symbol in Figure 29-15, we get an N-channel MOSFET of the enhancement type.

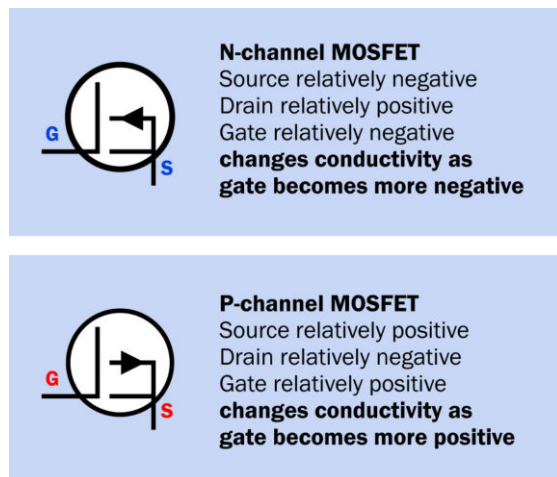


Figure 29-14. Either of the two symbols can be combined with either of the two symbols in the next figure, to create one of the four symbols for a MOSFET. See text for details.

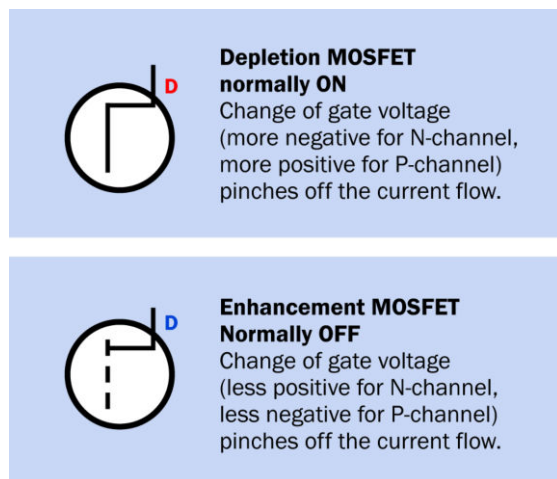


Figure 29-15. Either of the two symbols can be combined with either of the two symbols from the previous figure, to create one of the four symbols for a MOSFET. See text for details.

In an additional attempt to clarify MOSFET behavior, four graphs are provided in [Figure 29-16](#), [Figure 29-17](#), [Figure 29-18](#), and [Figure 29-19](#). Like JFETs, MOSFETs have an initial ohmic region, followed by a saturation region where current flows relatively freely through the device. The gate-to-

source voltage will determine how much flow is permitted. However, it is important to pay close attention to the graph scales, which differ for each of the four types of MOSFET.

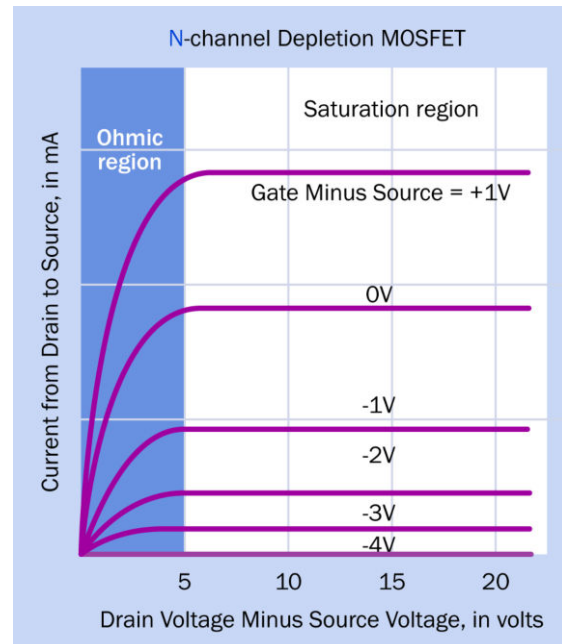


Figure 29-16. Current flow through a depletion-type, N-channel MOSFET.

In all of these graphs, a bias voltage exists, which allows zero current to flow (represented by the graph line superimposed on the horizontal axis). In other words, the MOSFET can operate as a switch. The actual voltages where this occurs will vary with the particular component under consideration.

The N-channel, enhancement-type MOSFET is especially useful as a switch because in its normally-off state (with zero bias) it presents a very high resistance to current flow. It requires a relatively low positive voltage at the gate, and effectively no gate current, to begin conducting conventional current from its drain terminal to its source terminal. Thus it can be driven directly by typical 5-volt logic chips.

Depletion-type MOSFETs are now less commonly used than the enhancement-type.

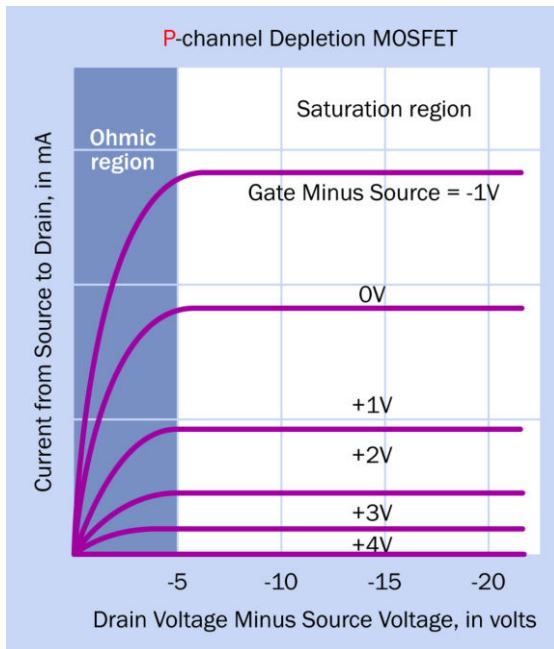


Figure 29-17. Current flow through a depletion-type, P-channel MOSFET.

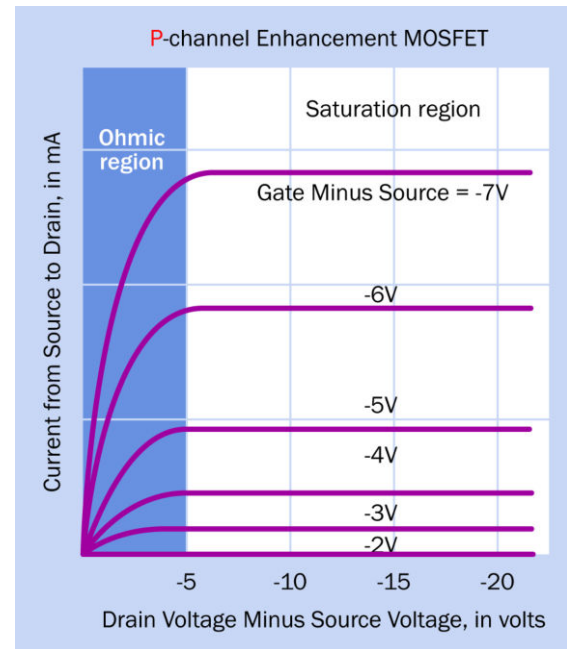


Figure 29-19. Current flow through an enhancement-type, P-channel MOSFET.

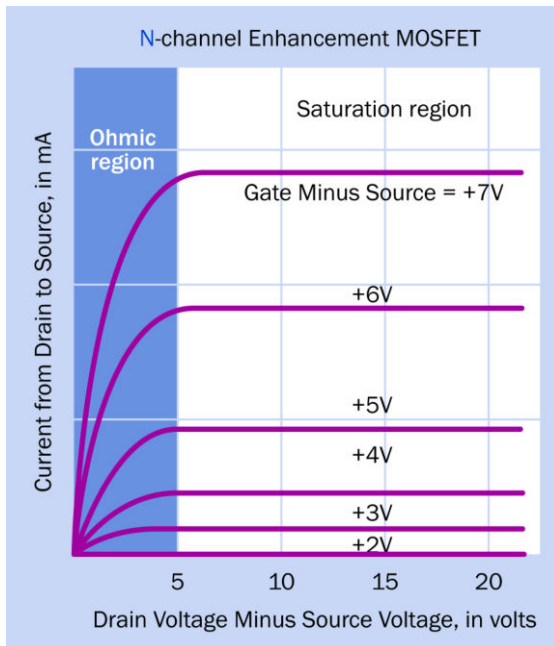


Figure 29-18. Current flow through an enhancement-type, N-channel MOSFET.

The Substrate Connection

Up to this point, nothing has been said about a fourth connection available on many MOSFETs, known as the *body terminal*. This is connected to the substrate on which the rest of the component is mounted, and acts as a diode junction with the channel. It is typically shorted to the source terminal, and in fact this is indicated by the schematic symbols that have been used so far. It is possible, however, to use the body terminal to shift the threshold gate voltage of the MOSFET, either by making the body terminal more negative than the source terminal (in an N-channel MOSFET) or more positive (in a P-channel MOSFET). Variants of the MOSFET schematic symbols showing the body terminal are shown in [Figure 29-20](#) (for depletion MOSFETs) and [Figure 29-21](#) (for enhancement MOSFETs).

A detailed discussion of the use of the body terminal to adjust characteristics of the gate is beyond the scope of this encyclopedia.

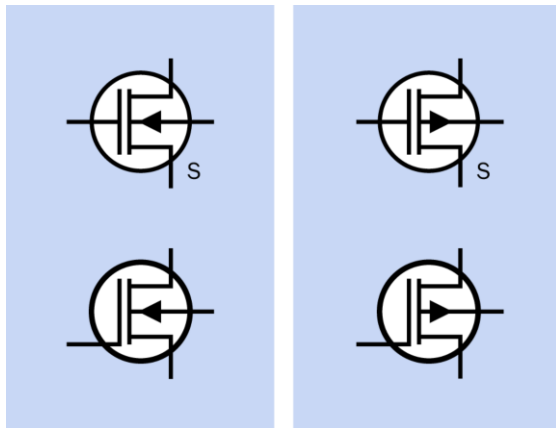


Figure 29-20. Schematic symbol variants for depletion MOSFETs, showing the body terminal separately accessible instead of being tied to the source terminal.

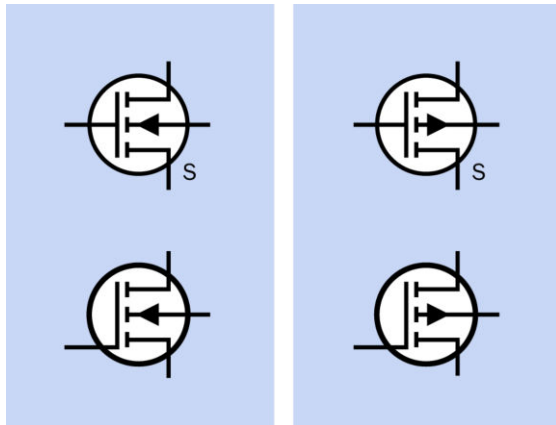


Figure 29-21. Schematic symbol variants for enhancement MOSFETs, showing the body terminal separately accessible instead of being tied to the source terminal.

Variants

A few FET variants exist in addition to the two previously discussed.

MESFET

The acronym stands for *MEtal-Semiconductor Field Effect Transistor*. This FET variant is fabricated from gallium arsenide and is used primarily in radio frequency amplification, which is outside the scope of this encyclopedia.

V-Channel MOSFET

Whereas most FET devices are capable of handling only small currents, the *V-channel MOSFET* (which is often abbreviated as a *VMOS FET* and has a V-shaped channel as its name implies) is capable of sustained currents of at least 50A and voltages as high as 1,000V. It is able to pass the high current because its channel resistance is well under 1Ω. These devices, commonly referred to as *power MOSFETs*, are available from all primary semiconductor manufacturers and are commonly used in switching power supplies.

Trench MOS

The TrenchMOS or Trenchgate MOS is a MOSFET variant that encourages current to flow vertically rather than horizontally, and includes other innovations that enable an even lower channel resistance, allowing high currents with minimal heat generation. This device is finding applications in the automobile industry as a replacement for electromechanical **relays**.

Values

The maximum values for JFETs, commonly found listed in datasheets, will specify V_{ds} (the drain-source voltage, meaning the potential difference between drain and source); V_{dg} (the drain-gate voltage, meaning the potential difference between drain and gate); V_{gsr} (the reverse gate-source voltage); gate current; and total device dissipation in mW. Note that the voltage differences are relative, not absolute. Thus a voltage of 50V on the drain and 25V on the source might be acceptable in a component with a V_{ds} of 25V. Similarly, while a JFET's "pinch-off" effect begins as the gate becomes "more negative" than the source, this can be achieved if, for example, the source has a potential of 6V and the gate has a potential of 3V.

JFETs and MOSFETs designed for low-current switching applications have a typical channel resistance of just a few ohms, and a maximum switching speed around 10Mhz.

The datasheet for a MOSFET will typically include values such as gate threshold voltage, which may be abbreviated V_{gs} (or V_{th}) and establishes the relative voltage at which the gate starts to play an active role; and the maximum on-state drain current, which may be abbreviated $I_{d(on)}$ and establishes the maximum limiting current (usually at 25 degree Centigrade) between source and gate.

How to Use it

The combination of a very high gate impedance, very low noise, very low quiescent power consumption in its off state, and very fast switching capability makes the MOSFET suitable for many applications.

P-Channel Disadvantage

P-channel MOSFETs are generally less popular than N-channel MOSFETs because of the higher resistivity of P-type silicon, resulting from its lower carrier mobility, putting it at a relative disadvantage.

Bipolar Substitution

In many instances, an appropriate enhancement-type MOSFET can be substituted for a bipolar transistor with better results (lower noise, faster action, much higher impedance, and probably less power dissipation).

Amplifier Front Ends

While MOSFETs are well-suited for use in the front end of an audio amplifier, chips containing MOSFETs are now available for this specific purpose.

Voltage-Controlled Resistor

A simple voltage-controlled resistor can be built around a JFET or MOSFET, so long as its performance remains limited to the linear or ohmic region.

Compatibility with Digital Devices

A JFET may commonly use power supplies in the range of 25VDC. However, it can accept the high/

low output from a 5V digital device to control its gate. A 4.7K pullup resistor is an appropriate value to be used if the FET is to be used in conjunction with a TTL digital chip that may have a voltage swing of only approximately 2.5V between its low and high thresholds.

What Can Go Wrong

Static Electricity

Because the gate of a MOSFET is insulated from the rest of the component, and functions much like a plate of a capacitor, it is especially likely to accumulate static electricity. This static charge may then discharge itself into the body of the component, destroying it. A MOSFET is particularly vulnerable to electrostatic discharge because its oxide layer is so thin. Special care should be taken either when handling the component, or when it is in use. Always touch a grounded object or wear a grounded wrist band when handling MOSFETs, and be sure that any circuit using MOSFETs includes appropriate protection from static and voltage spikes.

A MOSFET should not be inserted or removed while the circuit in which it performs is switched on or contains residual voltage from undischarged capacitors.

Heat

Failure because of overheating is of special concern when using power MOSFETs. A Vishay Application Note ("Current Power Rating of Power Semiconductors") suggests that this kind of component is unlikely to operate at less than 90 degrees Centigrade in real-world conditions, yet the power handling capability listed in a datasheet usually assumes an industry standard of 25 degrees Centigrade.

On the other hand, ratings for continuous power are of little relevance to switching devices that have duty cycles well below 100%. Other factors also play a part, such as the possibility of power surges, the switching frequency, and the integrity of the connection between the component

and its heat sink. The heat sink itself creates uncertainty by tending to average the temperature of the component, and of course there is no simple way to know the actual junction temperature, moment by moment, inside a MOSFET.

Bearing in mind the accumulation of unknown factors, power MOSFETs should be chosen on an extremely conservative basis. According to a tutorial in the *EE Times*, actual current switched by a MOSFET should be less than half of its rated current at 25 degrees, while one-fourth to one-third are common. [Figure 29-22](#) shows the real-world recommended maximum drain current at various temperatures. Exceeding this recommendation can create additional heat, which cannot be dissipated, leading to further accumulation of heat, and a thermal runaway condition, causing eventual failure of the component.

Wrong Bias

As previously noted, applying forward bias to a JFET can result in the junction between the gate and the source starting to behave like a forward-biased diode, when the voltage at the gate is greater than the voltage at the source by approximately 0.6V or more (in an N-channel JFET). The junction will present relatively little resistance, encouraging excessive current and destructive consequences. It is important to design devices that allow user input in such a way that user error can never result in this eventuality.

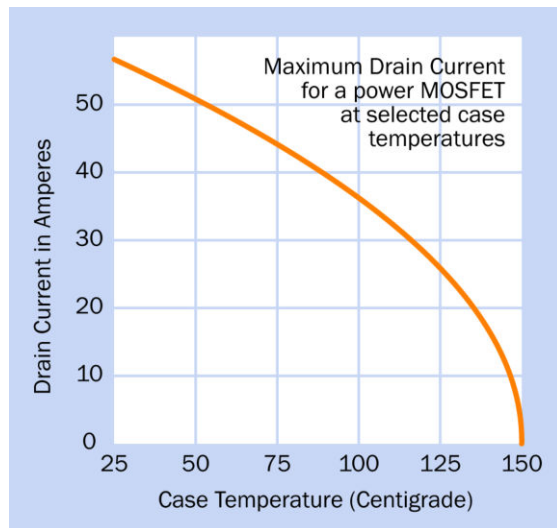


Figure 29-22. Maximum advised drain current through a power MOSFET, related to case temperature of the component. Derived from *EE Times Power MOSFET Tutorial*.



Schematic Symbols

This section contains a compilation of schematic symbols for components that have been described in this volume. They are sequenced primarily in alphabetical order, as this section is intended for use as an index. However, symbols that have a strong similarity are grouped together; thus **potentiometer** is found adjacent to **resistor**, and all types of transistors are in the same group.

The symbol variants shown in each blue rectangle are functionally identical.

Where a component has mandatory polarity or is commonly used with a certain polarity, a red plus sign (+) has been added for guidance. This sign is not part of the symbol. In the case of polarized capacitors, where a plus sign is normally shown (or should be shown) with the symbol, the plus sign is a part of the symbol and appears in black.

This is not intended to be an exhaustive compilation of symbol variants. Some uncommon ones may not be here. However, the list should be sufficient to enable identification of components in this volume.

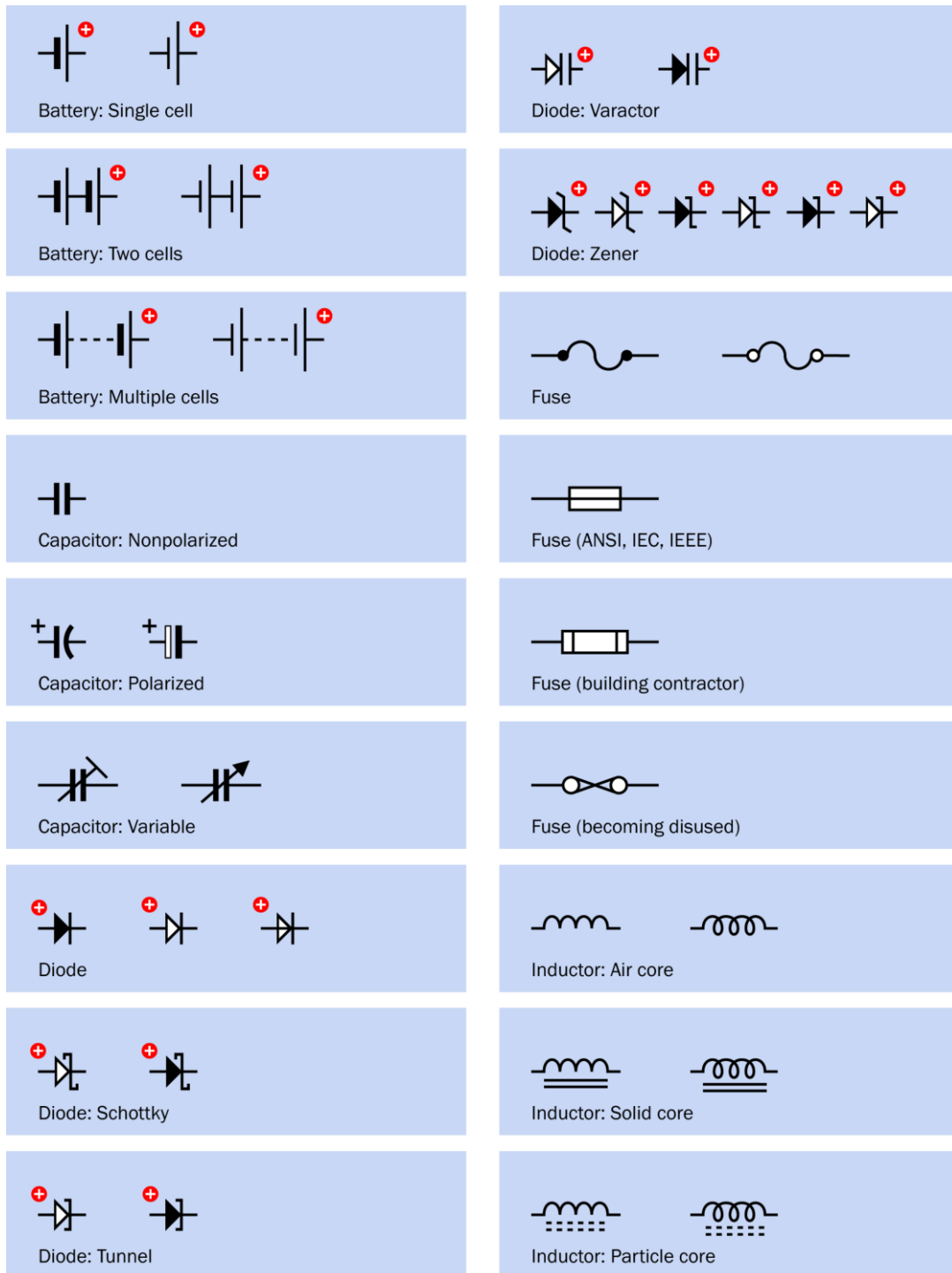


Figure A-1. Schematic symbols

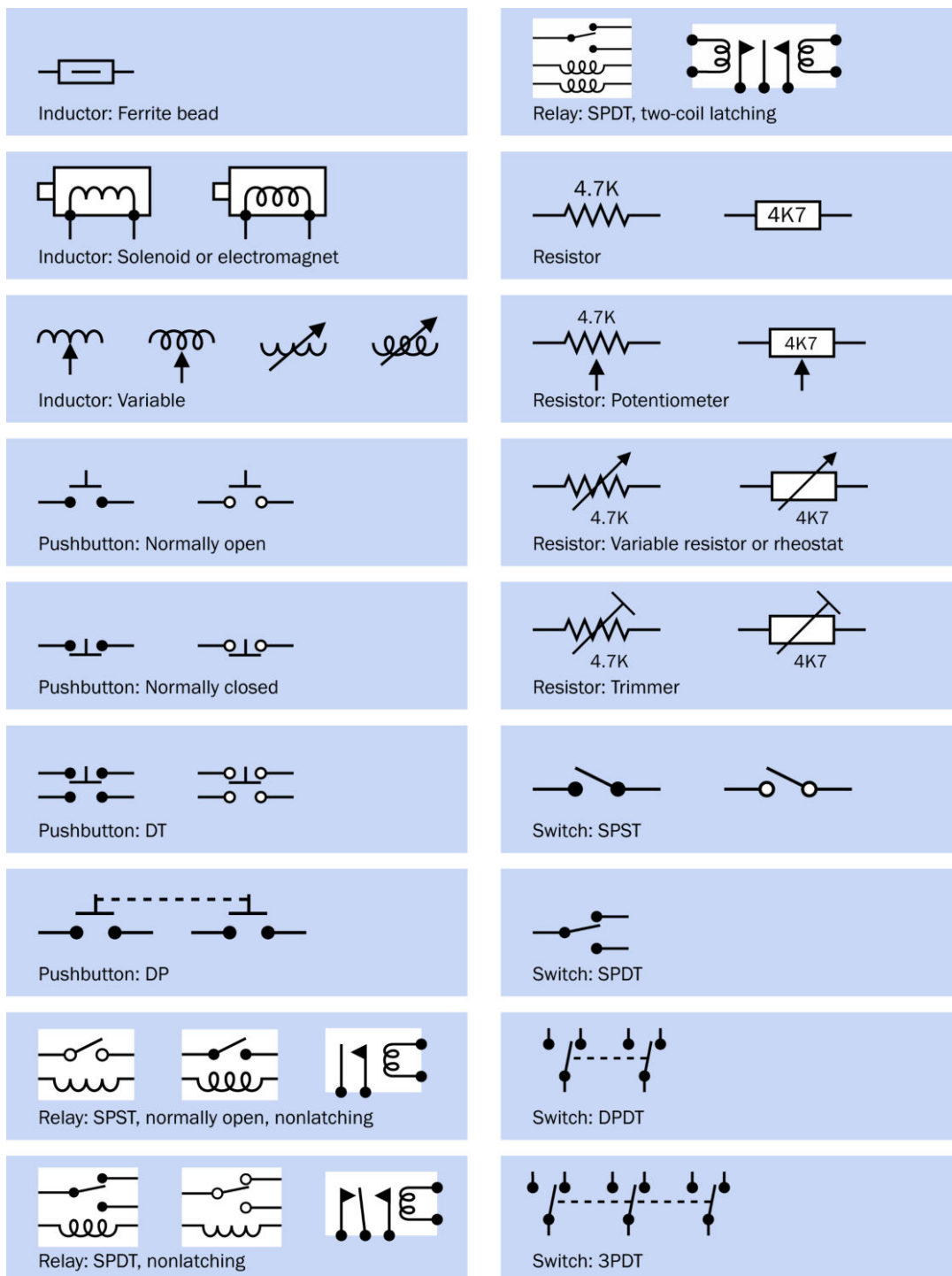


Figure A-2. Schematic symbols, continued

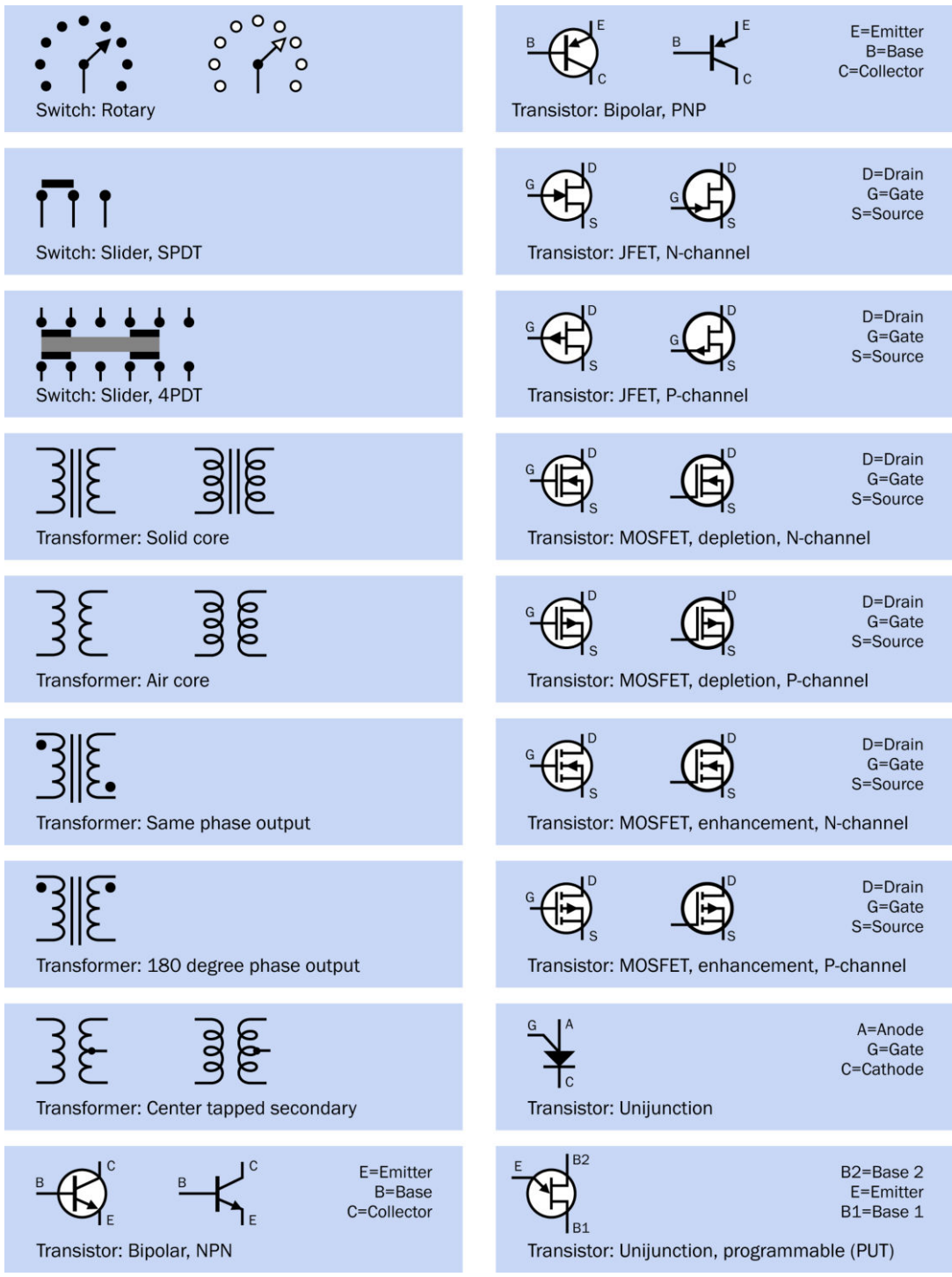


Figure A-3. Schematic symbols, continued

Symbols

18 * D) + (40 * W, 128

1:1 transformers (see isolation transformers)

A

absolute encoding, 53

AC adapters, 142, 143, 157, 159

AC inrush, 177

AC motors, 191–200

 design of, 191

 function of, 191

 potential problems, 200

 types of, 195

 use of, 199

 values for, 199

AC-AC transformers, 135–142

 design of, 136

 function of, 135

 potential problems, 142

 schematic symbol for, 135

 types of, 138

 use of, 142

 values for, 141

AC-DC power supplies, 143–147

 function of, 143

 potential problems, 147

 types of, 143

 use of, 147

accumulators, 5

active mode, 245

actuators, 38

adjustable inductors, 125

adjustable output converters, 152

adjustable voltage regulators, 163

Ah/AH/A/H (see amp-hours)

air cores, 125

alkaline cells, 8, 13

alternate pushbuttons, 33

alternating current (AC), 135, 143, 157, 196, 221

amp-hour capacity, 11, 13,

amp-hours (Ah), 11

amplifiers, 250

amplitude-modulated, 125

analog servo motors, 204

annulus, 183

anodes, 6, 221

anti-backlash gears, 114

apparent power, 141

arcing, 47, 72, 108

armatures, 66

audio tone control, 85

audio transformers, 140

audio-taper potentiometers, 91, 96

automotive fuses, 23

automotive relays, 69

autotransformers, 140

auxilliary winding, 196

avalanche breakdown, 233

axial leads, 99

axial loading, 185

axial resistors, 76

B

back EMF, 108, 218, 227, 228

backlash, 188, 204

bandpass filters, 131

basic switches, 39

batteries, 5–16

 capacitors as, 109

 capacity, 11

 DC-AC inverters and, 160

- design of, 6
- disposable, 7, 8
- fuel cells, 8
- function of, 5
- potential problems, 15
- rechargeable, 7, 9
- schematic symbols for, 6
- use of, 14
- values for, 11
- voltage, 13
- battery acid, 9
- battery chargers, 8
- bearings
 - failure of, 189
- beta value (β), 244
- biased relays, 67
- biased semiconductors, 241
- bifilar motors, 216
- bipolar electrolytic capacitors, 102
- bipolar junction transistor (BJT) (see bipolar transistors)
- bipolar motors, 211
- bipolar stepper motors, 213
- bipolar transistors, 241–252
 - design of, 241
 - function of, 241
 - potential problems, 251
 - schematic symbols for, 242
 - types of, 245
 - use of, 246
- BLDC motors (see brushless DC motors)
- blown fuses, 21, 27
- body terminals, 261
- boost-type converters, 151, 154
- boots, 43
- break-before-make switches, 39, 53
- breakdown state, 256
- breakdown voltage, 224
- breakers, 160
- breaking capacity, 22
- bridge rectifiers, 226
- brushes, 179, 187, 198
- brushless DC motors, 183
- buck converters, 150
- buck-boost converters (see fly-back converters)

- bulk metal foil, 83
- bushings, 38
- button capacitors (see disk capacitors)
- button cells, 5
- bypass capacitors, 107, 154
- bypass switches, 176

C

- c-clamps, 110
- C-rate, 12
- capacitance, 105, 115
- capacitive reactance, 106
- capacitor arrays, 101
- capacitor clamps, 110
- capacitor networks, 101
- capacitors, 8, 97–111, 97
 - (see also variable capacitors)
 - as alternative to shorting coils, 195
 - design of, 97
 - function of, 97
 - potential problems, 109
 - schematic symbols, 97
 - types of, 99
 - use of, 107
 - values for, 104
- capacity, 11–14
- caps (see capacitors)
- carbon-composite resistors, 82
- carbon-film potentiometers, 95
- carbon-film resistors, 82
- carriers, 183
- cartridge fuses, 23
- cathode-ray tubes, 171
- cathodes, 6, 221
- cells, 5, 6
- ceramic capacitors, 103
- charging
 - incorrect voltage and, 15
 - reverse charging, 16
- chatter, 72
- chokes, 119
- chopper drives, 217
- circuit breakers, 15, 21
- circuits
 - RC circuits, 79
- clamp diodes, 230
- closed-loop systems, 210, 217
- CMOS (complementary metal oxide semiconductor), 253
- code samples
 - digital access to, 3
- coded rotary switches, 53
- cogging, 193
- coils, 66, 121, 135, 170
- cold cranking rating, 159
- collectors, 242, 246
- common contacts, 38
- common-bus configurations, 78
- common-collector amplifiers, 250
- common-emitter amplifier, 250
- commutators, 179, 195, 198
- complement-coded rotary switches, 54,
- complementary metal oxide semiconductor (see CMOS)
- components
 - definition of, 2
- condensers (see capacitors)
- contact bounce, 61
- contacts
 - burnout of, 63
 - contamination of, 48, 73
 - overloading of, 57
 - relays and, 66
 - vulnerability of, 57
- continuous inductors, 122
- contractors, 70
- control shaft encoders, 59
- conventional current, 7
 - inductors and, 120
- converters (see DC-DC converters)
- cores, 119, 132, 137, 138
- coupling capacitors, 107
- covered power supplies, 146
- crossover networks, 130
- crystal diodes (see diodes)
- crystal sets, 125
- cup (potentiometer), 93
- current, 5
 - inadequate, 15
 - stall current, 184
- current amplifiers, 243, 253
- current gain (hFE), 244

- current ratings
 - DC motors, 184
 - fuses, 22
 - pushbuttons, 35
 - relays, 70
 - rotary switches, 56
 - rotational encoders, 61
- cutoff mode, 245
- cylindrical capacitors, 99

D

- damping, 220
- Darlington chips, 249
- Darlington pair chips, 251
- Darlington pairs, 161, 218, 241, 248
- DC current
 - AC-AC transformers and, 142
 - inductors and, 121
- DC motors, 179–189
 - current rating, 184
 - design of, 179
 - function of, 179
 - potential problems, 187
 - types of, 181
 - use of, 185
- DC-AC inverters, 157–160
 - design of, 157
 - function of, 157
 - potential problems, 160
 - types of, 158
 - use of, 159
 - values for, 158
- DC-DC converters, 149–155
 - design of, 149
 - function of, 149
 - potential problems, 155
 - types of, 150
 - use of, 154
 - values for, 152
- DCR (DC resistance), 129
- debounce, 48
- decks, 52
- decoupling capacitors, 107
- deep cycle batteries, 9
- degauss, 171
- delay fuses, 23
- delta configuration, 181
- demagnetization
 - of rotors, 220
- demagnetizing objects, 171
- depletion device, 255
- depletion layer, 255
- depletion region, 223
- detent torque, 220
- detents, 59
- dielectric absorption, 110
- dielectric constant, 105
- dielectric memory, 110
- dielectrics, 97, 103, 113, 257
- digital access
 - code samples, 3
- digital servo motors, 204
- diode arrays, 226
- diodes, 221–233
 - DC circuits and, 47
 - design of, 223
 - function of, 221
 - potential problems, 232
 - schematic symbol for, 221
 - types of, 224
 - use of, 227
 - values for, 226
- DIP (see dual-inline package)
- DIP switches, 43, 53
 - (see also rotary DIP switches)
 - vs. jumpers, 17
- direct current (DC), 143, 157
- direct current-excited motors, 196
- discrete semiconductors, 241
- disk capacitors, 100
- disposable batteries, 7, 8–9
- doping, 253
- double-pole pushbuttons, 29
- double-pole switches, 37
- double-throw (DT) switches, 37
- drain-source breakdown, 256
- dropout voltage, 161
- dry cells, 7
- dry joints, 48
- dual-coil latching relays, 67
- dual-ganged potentiometers, 93
- dual-inline package (DIP), 77
- dynamic resistance, 224, 225
- dyne-meters, 185

E

- effective series resistance (ESR), 98, 154
- efficiency
 - converters and, 153
- 8-wire motors, 216
- electrical capacity (see capacity)
- electrical noise, 87, 97, 132, 147, 155, 207, 230
- electrical polarity, 123
- electrochemical cells, 5
- electrochemical power sources, 5
- electrodes, 6, 7
- electrolysis, 5
- electrolytes, 6
- electrolytic capacitors, 101, 110, 147
- electrolytic cells, 5
- electrolytic layer dielectrics, 103
- electromagnetic armature relays, 65
- electromagnetic interference (EMI), 147, 154, 187
- electromagnets, 169–172
 - function of, 169, 169
 - potential problems, 172
 - types of, 170
 - use of, 171
 - values for, 171
- electromechanical linear actuators, 184
- electromechanical relays, 65
- electromotive force (EMF), 177, 218, 227, 228
- electron holes, 98, 223
- electronic commutation, 184
- electrostatic discharge (ESD), 225, 263
- element, 21
- emergency switches, 35
- EMI (see electromagnetic interference)
- emitters, 242, 246
- enable pins, 165
- encoded output rotary switches, 53
- enhancement devices, 255, 258

epicyclic gears, 182
equivalent series resistance (ESR), 110
errata, web address for, 4
ESD (see electrostatic discharge)

F

farads (F), 104
ferrite beads, 126
ferromagnetic cores, 137
field effect transistor (FET)
 design of, 253
 function of, 253
 potential problems, 263
 types of, 262
 use of, 263
 values for, 262
field windings, 192
field-effect transistor (FET), 253–264
555 timer, 186, 202, 214, 218, 235
flat-ganged potentiometers, 93
flip-flop behavior, 67
flooded design, 9
flux density, 170
flyback converters, 151
flyback diodes, 47
flyback-type converters, 144
foot pedal pushbuttons, 33
FORM A/B switches, 39
Form A/B/C pushbuttons, 30
formulae
 presentation of, 3
forward bias, 241
forward direction (diodes), 221
forward EMF, 177
frames, 173
freewheeling diodes, 47
front panel mount switches, 45
fuel cells, 8
fuseholders, 27
fuses, 15, 21–28
 (see also circuit breakers)
 current rating, 22
 design of, 21
 function of, 21
 potential problems, 27

 schematic symbols for, 21
 types of, 22
 use of, 26, 28
 voltage rating, 22
fusible links, 24
fusible resistors, 82

G

ganged potentiometers, 93
gear motors (see gearhead motors)
gearhead motors, 181, 188
gears
 nylon vs. metal, 204
general-purpose resistors, 76
generators, 120
gigaohms, 79
Gray-coded outputs, 55
grounding
 common ground, 142
 variable capacitors and, 117
Gunn diodes, 221, 226
gyrators, 127, 132

H

H bridge system, 187
half-cells, 6
half-stepping mode, 217
Hall effect sensors, 184
harmonics, 157
header sockets, 18
headers, 18, 34
heat
 AC motors and, 200
 capacitors and, 110
 converters and, 155
 diodes and, 233
 effect on motors, 188
 power resistors and, 84
 resistors and, 87
 solenoids and, 177
 voltage regulators and, 166
heat sinks, 76, 246
Henry units, 128
high phase count motors, 216

high side, 46
high-frequency transistors, 245
high-pass filters, 107, 130
high-speed diodes, 224
hold current, 25
holding torque, 220
horse-shoe magnet design, 170
hot switching, 37
hunting, 220
hybrid motors, 216
hysteresis, 122, 124, 170, 220
hysteresis motors, 197

I

IGBT (see insulated-gate bipolar transistor)
illuminated pushbuttons, 31
impedance, 140
in-line fuse holders, 24,
inaccuracy
 resistors and, 87
incandescent bulbs, 31
incremental devices
 rotational encoders, 59
indicators, 84
inductance, 87, 128
inductance index, 129
inductance-capacitance circuits (see LC circuits)
induction motors, 210
induction rotors, 192
inductive loads
 switch choices for, 48
 voltage spikes and, 71
inductive reactance, 98, 106
inductors, 119–133
 design of, 120
 function of, 119
 potential problems, 132
 schematic symbol for, 119
 types of, 124
 use of, 130
 values for, 128
inrunner configuration, 184
insulated-gate bipolar transistor (IGBT), 186

insulated-gate field-effect transistor (IGFET), 253
integrated circuits, 241
internal resistance, 11
inverted AC motors, 199
ions, 6
ISAT (saturation current, 129
isolated vs. non-isolated, 154
isolation transformers, 140

J

jumper assemblies, 18
jumper sockets (see jumpers)
jumper wires, 17
jumpers, 17–19
junction field-effect transistor (JFET), 253, 253–256
junction threshold voltage, 223, 236

K

Karbonite gears, 204
keylock switches, 56
keypad pushbuttons, 34
kilohms, 79
knife switches, 37

L

latching pushbuttons, 33
latching relays, 46, 67
LC (inductance-capacitance) circuits, 115
lead-acid batteries, 5, 9, 15
leakage, 98, 223, 241, 252
LED (light-emitting diode), 36, 221
 pushbuttons and, 31
 resistors and, 84
limit switches, 39, 184, 187
linear actuators, 184, 186, 218
linear mode, 245
linear motors, 195

linear regulated power supplies, 143
linear taper potentiometers, 90
linear voltage regulators (see voltage regulators)
lithium-ion batteries (Li-ion), 9, 15
lithium-ion polymer batteries, 9
load regulation, 153
loads, 5
lock paint, 117
locked-rotor torque, 194
locking pushbuttons, 33
locking toggle switches, 43
log-taper potentiometers, 91, 96
logarithmic volume control, 95,
logic chips, 48, 154
logic circuits, 47, 157
logic gates, 230
logic inverters, 157
low-dropout voltage regulators, 161, 164
low-pass filters, 108, 130
low-signal relays, 69
lubricants
 environmental effects on, 188

M

magnetic circuits, 121
magnetic cores, 122, 124
magnetic domains, 122, 132, 170
magnetic fields, 72
magnetic polarity, 123
magnetic resistance (see reluctance)
magnetic saturation, 122
mAh (see milliamp-hours)
MAKE magazine
 contact information, 4
make-before-break switches, 39, 53
make-to-break connections, 30, 39
make-to-make connections, 30
matrix encoded pushbuttons, 34
mechanical encoders, 55
mechanical rotary encoders, 59

mechanical wear
 of switches, 48
 potentiometers, 95
megohms, 79
membrane pad pushbuttons, 34
membrane separators, 6
memory effect, 10
MESFET (MEtal-Semiconductor Field Effect Transistor), 262
metal-film resistors, 83
metal-oxide semiconductor field-effect transistor (see MOSFET)
meter loading, 88
microcontrollers
 as alternative to switches, 47
 as pulse sources, 186
 increasing output of, 248
 keypads and, 34
 pushbuttons and, 33
 rotary switches and, 56
 rotational encoders and, 59, 62
 servo motors and, 202
microfarads (μF), 104, 111
microhenrys (μH), 128
microstepping mode, 217
microswitches, 39
milliamp-hours (mAh), 11
millihenrys (mH), 128
milliohms, 79
miniaturized potentiometers, 92
modified sine waves, 158
monetary switches, 29, 67
MOSFET (metal-oxide semiconductor field-effect transistor), 149, 186, 253, 256–261, 263
movable contacts, 38
multiphase motors, 217
multiple-turn potentiometers, 93
Must Operate By voltage, 70
Must Release By voltage, 70
mutual induction, 136

N

N layer, 223
N-channel, 253
N-type semiconductors, 241
nanofarads (nF), 104

- nanohenrys (nH), 128
- negative resistance region, 237
- negative voltage regulators, 164
- neon bulbs, 31
- neutral relays, 67
- Newton-meters, 185
- nichrome (Ni-chrome) elements, 77
- nickel-cadmium batteries (nicad/ NiCd), 9, 10
- nickel-metal hydride batteries (NiMH), 9, 10
- noise
 - electrical, 87, 97, 132, 147, 155, 207
 - reduction of, 189, 230
- non-excited synchronous motors, 196
- non-flammable resistors, 82
- nonlatching relays, 67
- nonmagnetic cores, 125
- nonshorting switches, 53, 57
- normally closed (NC), 31, 39
- normally open (NO), 31, 39
- notch filters, 131
- nuisance opening, 27
- nylon vs. metal gears, 204

O

- octal base, 69
- OCV (see open circuit voltage)
- ohmic region, 256
- on-off switches, 37
- open circuit voltage (OCV/Voc), 13
- open-frame power supplies, 146
- open-frame switches, 51
- open-loop systems, 210, 217
- operating point, 250
- optical mouse, 59
- oscillating, 220
- ounce-inches, 185
- output
 - voltage regulators and, 166
- outrunner configuration, 184
- overcoat material, 117

- overheating
 - DC-AC inverters and, 160
 - MOSFETs, 263
 - potentiometers, 96
 - transformers and, 142
- overshoot, 219

P

- P layer, 223
- P-channel, 253
- P-type semiconductors, 241
- paddle switches, 44
- panel-mount switches, 45, 49
- panel-mounted pushbuttons, 32
- parallel batteries
 - high current flow, 16
- parasitic capacitance, 132
- parasitic inductance, 87, 103
- pass transistors, 161
- PC board rotary switches, 55
- PC mount switches, 45
- PC pushbuttons, 32
- PC terminals, 45
- peak inrush, 147, 177
- peak inrush current, 27
- permanent magnet stepper motors, 211
- permeability, 122
- Peukert number, 12
- phase count, 209
- phases (steps), 209
- pick-offs, 90
- picofarads (pF), 104
- piles, 5
- pin configurations
 - correct identification of, 167
 - relays, 67, 72
- PIN diodes, 221, 226
- pitch, 18
- planetary gears, 182
- plastic-film dielectrics, 103
- plug fuses, 22
- plug-in transformers, 139
- plungers, 173
- PN junction diodes (see diodes)
- polarity
 - danger of incorrect, 15, 110
 - danger of reversed, 233
 - DC relays and, 67
 - mnemonics for, 123
 - preventing reversed, 142
- polarized relays, 67
- pole faces, 173
- poles, 29, 30, 37, 38, 38
- polymeric positive temperature coefficient fuse, 24
- polyphase AC, 196
- positive charges, 98
- positive voltage regulators, 164
- potentiometers, 75, 89–96, 186
 - design of, 90
 - function of, 89
 - potential problems, 95
 - schematic symbols for, 89
 - types of, 90
 - use of, 94
- pound-feet, 185
- power cells, 5
- power cords
 - correct choice of, 142
- power MOSFETs, 262
- power resistors, 76
- power transformers, 139
- power transistors, 245
- power wire-wound resistors, 83
- PPR (see pulses per rotation)
- precision resistors, 76
- precision wire-wound resistors, 83
- press-twice pushbuttons, 33
- primary cells, 7
- primary coils, 135
- programmable unijunction transistor (PUT), 235, 239
- protection diodes, 218, 228
- proximity effect, 133
- PTC/PPTC (see polymeric positive temperature coefficient fuse)
- pull-in torque, 220
- pull-out torque, 220
- pulldown resistors, 85
- pullup resistors, 56, 85
- pulse-width modulation (PWM), 144, 150, 158, 186, 199, 201, 217
- pulses per rotation (PPR), 61

push-lock push-release pushbuttons, 33
 push-on push-off pushbuttons, 33
 push-push pushbuttons, 33
 pushbutton switches, 29
 pushbuttons, 29–36
 current ratings, 35
 design of, 29
 function of, 29
 potential problems, 35
 schematic symbol for, 29
 types of, 30
 use of, 35
 pushwheel switches, 55
 PUT (programmable unijunction transistor), 235, 239
 PWM (see pulse-width modulation)

Q

quadrature output, 61
 quality factor, 107
 quasi-low-dropout voltage regulators, 165
 quick connect terminals, 45
 quiescent current, 166
 quiescent point, 250

R

radial leads, 99
 radial loading, 185
 radial resistors, 76
 radio buttons, 35
 radio crystal set, 125
 radio frequencies (RF), 133
 ramping, 219
 rated current
 fuses, 22
 rated voltage, 22
 RC (resistor-capacitor) network, 86, 98
 RC circuits, 79
 RC servo motors, 201
 reactance, 115, 123, 128

real-coded rotary switches, 54,
 rechargeable batteries, 7, 9–11
 rectified current, 199
 rectifier diodes, 224, 227, 233
 rectifiers, 142, 143, 144, 170, 221
 reed relays, 68, 72
 reed switches, 68, 171
 relative devices
 rotational encoders, 59
 relative encoding, 53
 relays, 65–73
 current rating, 70
 design of, 66
 function of, 65
 possible problems, 72
 schematic symbols for, 65, 67
 types of, 67
 use of, 71
 reluctance, 121, 129, 170, 176
 reluctance motors, 184, 197, 199
 reluctance stepper motors, 210
 remanent magnetism, 122
 reserve batteries, 9
 resettable fuses, 21, 24
 resistance, 210
 rotational encoders, 61
 resistor arrays, 77
 resistor ladders, 56
 resistor networks/ladders (see resistor arrays)
 resistor-capacitor (RC) network, 86, 98
 resistors, 75–88
 design of, 76
 function of, 75
 materials, 82
 possible problems, 87
 schematic symbols for, 75
 types of, 76
 use of, 84
 values for, 79
 resistors in parallel, 86
 resistors in series, 86
 resonant frequency, 106, 220
 response curves, 237
 retentivity, 124
 reverse charging, 16
 rheostats, 89, 186
 right-angle PC rotary switch, 54

right-angle PC terminals, 45
 right-hand rule, 123
 ripple-and-noise values, 154
 RMS (see root mean square)
 rocker switches, 40
 root mean square (RMS), 158
 rotary DIP switches, 53
 rotary switches, 51–57
 current rating, 56
 design of, 52
 function of, 51
 potential problems, 57
 schematic symbol for, 51
 types of, 52
 use of, 56
 rotational encoders, 47, 59–63
 current rating, 61
 design of, 59
 function of, 59
 potential problems, 62
 types of, 60
 use of, 62
 rotors, 51, 113, 179, 191, 192, 209
 demagnetization of, 220

S

Safari Books Online, 3
 safe operating area (SOA), 247
 saturated mode, 245
 saturation, 132, 220, 244
 saturation current, 129
 saturation region (I_{dss}), 256
 saturation, magnetic, 176
 schematic symbols, 265
 AC-AC transformers, 135, 138
 batteries, 6
 bipolar transistors, 242
 capacitors, 97
 diodes, 221
 fuses, 21
 inductors, 119
 MOFSET, 259
 potentiometers, 89
 pushbuttons, 29
 relays, 65, 67
 resistors, 75, 82
 rotary switches, 51

- slide switches, 41
- switches, 49
- unijunction transistors, 235
- Schottky diodes, 225
- SCR (see silicon-controlled rectifier)
- screw terminals, 45
- secondary cells, 7
- secondary coils, 135
- self-discharge rate, 8
- self-inductance, 119, 217
- self-resonance, 132
- self-resonant frequency, 106, 129
- semiconductors, 223
- series resistors, 84
- servo motors, 201–207
 - design of, 201
 - function of, 201
 - potential problems, 206
 - types of, 203
 - use of, 205
 - values for, 204
 - vs. stepper motors, 219
- settling time, 61
- shaded pole motors, 195
- shell cores, 138
- shielding
 - variable capacitors and, 117
- shock hazards
 - AC-AC transformers and, 142
 - AC-DC power supplies, 147
- short-circuited batteries, 15, 26
- short-circuited switches, 48
- shorting coils, 195
- shorting switches, 39, 53, 57
- shunts (see jumpers)
- signal clipping, 231
- signal diodes, 224, 233
- silicon-controlled rectifier (SCR), 186
- single side stable relays, 67
- single-coil latching relays, 67
- single-inline package (SIP), 77
- single-layer ceramic capacitors, 110
- single-phase induction motors, 195
- single-pole switches, 37
- single-throw (ST) switches, 37
- SIP (see single-inline package)
- SIP switches, 44
- skin effect, 133
- slide pushbuttons, 41
- slider potentiometers, 93
- slider pushbuttons, 31
- slider switches, 40, 48
- sliding noise, 62
- slip, 195
- Slo-Blo fuses, 23
- slow-blowing fuses, 26
- small cartridge fuses, 23
- small-signal relays, 69
- small-signal transistors, 245
- small-switching transistors, 245
- SMPS (see switching power supply)
- SMT (see surface-mount fuses)
- snap-action switches, 35, 39
- snap-in mount switches, 45
- snubbers, 35, 47, 79, 108
- SOA (see safe operating area)
- solder lug terminals, 45
- soldering damage, 27, 252
- solenoids, 173–177
 - design of, 174
 - function of, 173
 - potential problems, 177
 - types of, 176
 - use of, 177
 - values for, 176
- solid-state relays, 65
- splines, 204
- split-bobbin transformers, 141
- split-phase induction motors, 196
- sponge (lead), 9
- spur gears, 182
- squirrel-cage motors, 192
- squirrel-cage motors, 210
- SRF (self-resonant frequency, 129
- stability
 - resistors and, 82
- stall current, 184
- stall torque, 185
- standard type regulators, 161
- star configuration, 181
- stationary contacts, 38
- stators, 113, 179, 191, 209
- status pins, 165
- step angle, 210
- step loss, 219
- step motors (see stepper motors)
- step-down transformers, 136
- step-up transformers, 136
- stepper motors, 209–220
 - design of, 209
 - function of, 209
 - potential problems, 219
 - types of, 214
 - use of, 218
 - values for, 218
- stepping motors (see stepper motors)
- stops, 52
- strip fuses, 24
- subject-oriented organization, 2
- subminiature fuses, 24
- subminiature paddle switches, 44
- subminiature slide switches, 41
- subminiature switches, 40
- subpanel mount switches, 45
- substrate connections, 261
- sulfurization, 16
- sun gear, 183
- supercapacitors, 8, 109
- surface mount switches, 45
- surface-mount (SMT) fuses, 26
- surface-mount capacitors, 100
- surface-mount resistors, 77
- surface-mount transformers, 141
- surface-mounted pushbuttons, 32
- surge sensitivity, 22
- switch bounce, 48, 62
- switched potentiometers, 93
- switched-mode power supply (see switching power supply)
- switchers (see DC-DC converters) (see switching power supply)
- switches, 17, 37–49, 51
 - (see also rotary switches)
 - design of, 37
 - function of, 37
 - potential problems, 47
 - terminology, 38
 - types of, 38
 - use of, 46
 - values for, 45

switching diodes, 224
switching power supplies, 143, 144, 159
switching regulators (see DC-DC converters)
synchronous motors, 184, 196

T

tactile switches, 34
tank circuits (see LC circuits)
tantalum capacitors, 102, 110
taps, 137
Tcr/Tc (see temperature coefficient)
teeth (rotor), 210
temperature coefficient (Tcr/Tc), 82
terminals, choosing correct, 48
terminals, for switches, 45
THD (see total harmonic distortion)
thick-film resistors, 77, 83
thin-film resistors, 83
three-phase induction motors, 196
through-hole chips (capacitors), 101
through-hole fuses, 24, 26
through-hole terminals, 45
throws, 30, 38
thumbwheel switches, 55
time constant, 105, 130
time-delay relay switches, 70
toggle switches, 41
tolerance
 resistors, 79
toroidal core inductors, 126
torque
 expression of, 185
 fluctuations in, 193
 motor overloading and, 200
 stepper motors and, 220
total harmonic distortion (THD), 158
tracks, 90
transformer-based power supplies, 143

transient response, 166
transient suppressors, 228
transient voltage suppressor (TVS), 225
transients, 143
transistors, 241
 (see also bipolar transistors)
 resistors and, 84
transit time, 205
TrenchMOS, 262
triggering voltage, 236
trimmer capacitors, 113, 117
trimmer potentiometers, 89, 93, 113
Trimpots (see trimmer potentiometers)
trip current, 25
tripped fuses, 21
true sinewave inverters, 158
tuning capacitors, 113
tunnel diodes, 221, 226
turn rate, 205
TVS (see transient voltage suppressor)
two-contact pushbuttons, 29
two-pole switches, 37
two-way switches, 37

U

UJT (see unijunction transistor)
unijunction transistor (UJT), 235–240
 design of, 236
 function of, 235
 potential problems, 239
 schematic symbol for, 235
 types of, 238
 use of, 239, 240
 values for, 238
uninterruptible power supplies, 9
unipolar motors, 211, 214
universal motors, 198
universal stepper motors (see bifilar motors)
unloaded speed, 185

V

V-channel MOSFET (VMOS FET), 262
VA (see volts times amps)
vacuum tubes, 221, 241
valley voltage, 237
value coding, 88
 resistors, 81
valve-regulated lead-acid batteries (VRLA), 9
vandal resistant switches, 45
varactor diodes, 221, 225
variable capacitors, 113–117, 113
 (see also capacitors)
 design of, 113
 function of, 113
 possible problems, 117
 types of, 114
 use of, 115
 values for, 115
variable condensers (see variable capacitors)
variable frequency power supplies, 198
variable inductors, 125
variable reluctance, 210
variable resistors, 89
variable transformers, 140
variacs (see variable transformers)
varicaps (see varactor diodes)
Vc (see voltage coefficient)
Very Low Dropout voltage regulators, 165
vibration
 capacitors and, 110
 damping of, 220
 relays and, 73
Voc (see open circuit voltage)
voltage, 13
 automatic selection of, 229
 imbalance in AC motors, 200
 inaccurate in converters, 155
 inaccurate in voltage regulators, 167
 multiple outputs, 138,
 regulation with diodes, 230

- sensing with diodes, 232
- voltage amplifiers, 253
- voltage coefficient (Vc), 82
- voltage dividers, 56, 56, 75, 86, 163, 244
- voltage multipliers, 146
- voltage overload
 - bipolar transistors and, 252
 - capacitors and, 110
- voltage rating, 22
- voltage regulators, 161–167
 - design of, 161
 - function of, 161
 - potential problems, 166
 - types of, 163
 - use of, 165, 167
 - values for, 165

- voltage spikes
 - protection against, 71, 221
- voltage-controlled resistors, 256
- volts times amps, 141
- VRLA (see valve-regulated lead-acid batteries)

W

- wall-warts, 143
- windings, 121, 209
- wipers, 90, 140
- wire lead terminals, 45
- wire-wound resistors, 77
- wound-rotor AC induction motors, 198

- wye configuration, 181

Y

- Y configuration, 181
- yokes, 171

Z

- Zener diodes, 221, 225, 230, 231, 233
- Zener voltage, 230
- zero ohm components
 - resistors, 82
- zinc-carbon cells, 8

VOL. 2

Charles Platt with Fredrik Jansson

Encyclopedia of Electronic Components



Signal Processing

LEDs • LCDs • Audio • Thyristors
Digital Logic • Amplification



Make:
makezine.com

Encyclopedia of Electronic Components

Signal Processing

Want to know how to use an electronic component? This second book of a three-volume set includes key information on electronics parts for your projects—complete with photographs, schematics, and diagrams. You'll learn what each one does, how it works, why it's useful, and what variants exist. No matter how much you know about electronics, you'll find fascinating details you've never come across before.

Convenient, concise, well-organized, and precise

Perfect for teachers, hobbyists, engineers, and students of all ages, this reference puts reliable, fact-checked information right at your fingertips—whether you're refreshing your memory or exploring a component for the first time. Beginners will quickly grasp important concepts, and more experienced users will find the specific details their projects require.

- **Unique:** the first and only encyclopedia set on electronic components, distilled into three separate volumes
- **Incredibly detailed:** includes information distilled from hundreds of sources
- **Easy to browse:** parts are clearly organized by component type
- **Authoritative:** fact-checked by expert advisors to ensure that the information is both current and accurate
- **Reliable:** a more consistent source of information than online sources, product datasheets, and manufacturer's tutorials
- **Instructive:** each component description provides details about substitutions, common problems, and workarounds
- **Comprehensive:** Volume 1 covers power, electromagnetism, and discrete semiconductors; Volume 2 includes integrated circuits, and light and sound sources; Volume 3 covers a range of sensing devices.

Charles Platt

Charles Platt's lifelong love of electronics began when he built a telephone answering machine at age 15. A contributing editor to *Make:* magazine, he wrote the widely acclaimed *Make: Electronics*. He's also a science-fiction writer (author of *The Silicon Man*), and a former senior writer at *Wired* magazine.

US \$29.99 CAN \$31.99

ISBN: 978-1-4493-3418-5



Make:
makezine.com

Encyclopedia of Electronic Components Volume 2

Charles Platt
with Fredrik Jansson



Encyclopedia of Electronic Components Volume 2

by Charles Platt
with Fredrik Jansson

Copyright © 2015 Charles Platt. All rights reserved.

Printed in the United States of America.

Published by Maker Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

Maker Media books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (<http://safaribooksonline.com>). For more information, contact our corporate/institutional sales department: 800-998-9938 or corporate@oreilly.com.

Editor: Brian Jepson

Production Editor: Melanie Yarbrough

Proofreader: Jasmine Kwityn

Indexer: Last Look Editorial

Cover Designer: Karen Montgomery

Interior Designer: David Futato

Illustrator and Photographer: Charles Platt

November 2014: First Edition

Revision History for the First Edition:

2014-11-10: First release

See <http://oreilly.com/catalog/errata.csp?isbn=9781449334185> for release details.

Make:, Maker Shed, and Maker Faire are registered trademarks of Maker Media, Inc. The Maker Media logo is a trademark of Maker Media, Inc.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and Maker Media, Inc. was aware of a trademark claim, the designations have been printed in caps or initial caps.

While the publisher and the author have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the author disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

ISBN: 978-1-449-33418-5

[TI]

In fond memory of my father, Maurice Platt

Ahashare.com

Table of Contents

How to Use This Book..... xxi

> DISCRETE SEMICONDUCTOR

> > THYRISTOR

1. SCR..... 1

What It Does..... 1

How It Works..... 1

Switching Behavior..... 2

Internal Configuration..... 3

Breakdown and Breakover Voltage..... 4

SCR Concept Demo..... 4

AC Current Applications..... 5

Variants..... 5

Values..... 5

Commonly Used Abbreviations..... 6

How to Use It..... 6

Phase Control..... 7

Overvoltage Protection..... 7

What Can Go Wrong..... 8

Unexpected Triggering Caused by Heat..... 8

Unexpected Triggering Caused by Voltage..... 8

Confusion of AC and DC Ratings..... 8

Maximum Current versus Conduction Angle..... 9

Confusing Symbols..... 9

2. diac..... 11

What It Does..... 11

Symbol Variants	11
How It Works	12
Switching AC	13
Variants	13
Values	14
What Can Go Wrong	14
Unexpected Triggering Caused by Heat	14
Low-Temperature Effects	14
Manufacturing Tolerances	14
3. triac	15
What It Does	15
Symbol Variants	15
How It Works	17
Quadrants	17
Threshold, Latching, and Holding Current	18
Triac Testing	19
Breakover Voltage	20
Switching AC	20
Triac Triggered by a Diac	22
Other Triac Drivers	22
Charge Storage	23
Variants	23
Values	23
What Can Go Wrong	24
Unexpected Triggering Caused by Heat	24
Low-Temperature Effects	24
Wrong Type of Load	24
Wrongly Identified Terminals	24
Failure to Switch Off	24

> INTEGRATED CIRCUIT

> > ANALOG

4. solid-state relay	25
What It Does	25
Advantages	25
Disadvantages	26
How It Works	26
Variants	27
Instantaneous versus Zero Crossing	27
NC and NO Modes	28
Packaging	28
Solid-State Analog Switch	28
Values	29
How to Use It	29
What Can Go Wrong	30

Overheating Caused by Overloading	30
Overheating Caused by Bad Terminal Contact	30
Overheating Caused by Changing Duty Cycle	30
Overheating Caused by Component Crowding	30
Overheating in Dual Packaging	30
Reverse-Voltage Burnout	30
Low Voltage Output Current May Not Work	30
Inability to Measure AC Output	30
Relay Turns On but Won't Turn Off	31
Relays in Parallel Won't Work	31
Output Device Doesn't Run at Full Power	31
Solid-State Relays and Safety Disconnects	31

5. optocoupler 33

What It Does	33
How It Works	34
Variants	35
Internal Sensors	35
Basic Optocoupler Types	36
Values	36
How to Use It	36
What Can Go Wrong	37
Age	37
LED Burnout	37
Transistor Burnout	37

6. comparator 39

What It Does	39
Hysteresis	39
How It Works	39
Differences from an Op-Amp	42
Variants	42
Values	43
How to Use It	44
AND gate	45
Bistable Multivibrator	46
Relaxation Oscillator	46
Level Shifter	46
Window Comparator	46
Other Applications	47
What Can Go Wrong	47
Oscillating Output	47
Confused Inputs	48
Wrong Chip Type	48
Omitted Pullup Resistor	48
CMOS Issues	48
Erratic Output	48
Swapped Voltages	48
Heat-Dependent Hysteresis	48

7. op-amp	49
What It Does	49
How It Works	49
Dual Inputs	50
Negative Feedback	51
Op-Amps and Comparators	52
Variants	52
Values	52
How to Use It	53
Controlling the Gain	53
Calculating Amplification	53
Unintentional DC Voltage Amplification	54
Low-Pass Filter	54
High-Pass Filter	55
Relaxation Oscillator	55
Single Power Source	56
Offset Null Adjustment	56
What Can Go Wrong	57
Power Supply Problems	57
Bad Connection of Unused Sections	57
Oscillating Output	57
Confused Inputs	58
8. digital potentiometer	59
What It Does	59
Advantages	59
How It Works	60
Variants	61
Volatile and Nonvolatile Memory	61
Taper	62
Data Transfer	62
SPI	62
I2C Protocol	63
Up/Down Protocol	63
Other Control Systems	64
Connections and Modes	64
Values	65
How to Use It	66
Achieving Higher Resolution	66
What Can Go Wrong	67
Noise and Bad Inputs	67
Wrong Chip	67
Controller and Chip Out of Sync	67
Nonlinear Effects	67
Data Transfer Too Fast	67
9. timer	69
What It Does	69
Monostable Mode	69

Astable Mode	70
How It Works	70
Variants	70
The 555 Timer	70
555 Monostable Operation	71
555 Astable Operation	72
556 Timer	73
558 Timer	73
CMOS 555 Timer	74
5555 Timer	74
7555 Timer	74
7556 Timer	74
4047B Timer	75
Dual Monostable Timers	75
Values	76
555 Timer Values	76
Time Calculation in Monostable Mode	77
Time Calculation in Astable Mode	77
Dual Monostable Timers	77
How to Use It	79
555 Monostable Mode	79
555 Astable Mode	80
Separate Control of High and Low Output Times	80
555 Fifty Percent Astable Duty Cycle: 1	80
555 Fifty Percent Astable Duty Cycle: 2	81
Use of the 555 Control Pin	81
555 Flip-Flop Emulation	82
555 Hysteresis	83
555 and Coupling Capacitors	84
555 Loudspeaker Connection	84
Burst Mode	84
“You Lose” Game Sound	85
What Can Go Wrong	85
Dead Timer	85
CMOS Confused with Bipolar	86
The Pulse that Never Ends	86
Erratic Chip Behavior	86
Interference with Other Components	86
Erratic Behavior of Output Devices	86
Fatal Damage Caused by Inductive Loads	87

> > DIGITAL

10. logic gate	89
What It Does	89
Origins	89
How It Works	89
Inversion	90
Single-Input Gates	90

Gates with More than Two Inputs	91
Boolean Notation	91
Arithmetical Operations	91
Other Operations	92
Variants	93
Part Numbers	94
Families	95
Family Interoperability	96
Gates per Chip	96
Two Inputs, Single Gate	96
Three Inputs, Single Gate	97
Single Gate, Selectable Function	97
Two Inputs, Dual Gate	98
Original 74xx 14-Pin Format	98
Quad Two-Input 74xx Pinouts	98
Triple Three-Input 74xx Pinouts	99
Dual Four-Input 74xx Pinouts	100
Single Eight-Input 74xx Pinouts	100
74xx Inverters	101
Additional Variations	102
Pinouts in the Original 4000 Series	102
4000 Series Inverters	103
How to Use It	103
Which Family	103
Applications	104
What Can Go Wrong	105
Static	105
Floating Pins	105
Family Incompatibilities	105
Overloaded Outputs	105
Output Pulled Down	105
Incorrect Polarity and Voltages	105
Bent Pins	105
Unclean Input	106
Analog Input	106
11. flip-flop	107
What It Does	107
How It Works	107
NAND-Based SR Flip-Flop	108
NOR-Based SR Flip-Flop	109
Forbidden States	110
The JK Flip-Flop	112
Master-Slave Flip-Flop	113
D-Type Flip-Flops	114
Summary	116
Variants	116
Packaging	117
Values	117

How to Use It	118
What Can Go Wrong	119
Ambiguous Documentation	119
Faulty Triggering	119
Metastability	119
Other Issues	119
12. shift register	121
What It Does	121
Schematic Representation	122
How It Works	122
Abbreviations and Acronyms	123
Parallel Outputs and Inputs	123
Variants	124
Serial In, Serial Out	124
Serial In, Parallel Out	124
Parallel In, Serial Out	124
Parallel In, Parallel Out	125
Universal	125
Values	125
Power Considerations	126
Three-State Output	126
How to Use It	126
Dual Inputs	127
Preloading the Shift Register	127
Polling a Keyboard	127
Arithmetical Operations	127
Buffering	128
What Can Go Wrong	128
Confusing Classification	128
Inadequate Setup Time	128
Unconnected Input	129
Output Enable Issues	129
Floating Output Bus	129
13. counter	131
What It Does	131
Schematic Representation	131
How It Works	132
Modulus and Modulo	132
Pin Identifiers	133
Variants	134
Ripple versus Synchronous	134
Ring, Binary, and BCD	134
Clock Sources	135
Rising Edge and Falling Edge	136
Multiple Stages	136
Single and Dual	136
High-State, Low-State, and Three-State	136

Descending Output	136
Programmable Counters	137
Examples	137
Values	137
What Can Go Wrong	137
Lock-Out	137
Asynchronous Artifacts	137
Noise	138
14. encoder	139
What It Does	139
Schematic Symbol	139
Similar Devices	140
How It Works	140
Variants	141
Values	142
How to Use It	142
Cascaded Encoders	142
What Can Go Wrong	143
15. decoder	145
What it Does	145
Input Devices	145
LED Driver	146
Schematic Symbol	146
Similar Devices	146
How It Works	148
Variants	148
Values	148
How to Use It	149
What Can Go Wrong	149
Glitches	149
Unhelpful Classification	149
Active-Low and Active-High	149
16. multiplexer	151
What It Does	151
Differential Multiplexer	152
Similar Devices	152
How It Works	153
Schematic Symbol	153
Pin Identifiers	154
Variants	155
Values	155
How to Use It	156
Other Application Notes	157
What Can Go Wrong	157
Pullup Resistors	157
Break Before Make	157

Signal Distortion	157
Limits of CMOS Switching	157
Transients	157

> LIGHT SOURCE, INDICATOR, OR DISPLAY

> > REFLECTIVE

17. LCD	159
What It Does	159
How It Works	159
Variants	160
Active and Passive Types	161
Crystal Types	161
Seven-Segment Displays	161
Additional Segments	162
Dot-Matrix Displays	163
Color	166
Backlighting Options	167
Zero-Power Displays	167
How to Use It	167
Numeric Display Modules	167
Alphanumeric Display Module	168
What Can Go Wrong	169
Temperature Sensitivity	169
Excessive Multiplexing	170
DC Damage	170
Bad Communications Protocol	170
Wiring Errors	170

> > SINGLE SOURCE

18. incandescent lamp	171
What It Does	171
History	172
How It Works	172
Spectrum	173
Non-Incandescent Sources	174
Power Consumption	175
Variants	175
Miniature Lamps	175
Panel-Mount Indicator Lamps	176
Halogen or Quartz-Halogen	176
Oven Lamps	176
Base Variants	177
Values	177
Power	177
Illuminance	178

Intensity	178
MSCP	178
Efficacy	179
Efficiency	179
How to Use It	179
Relative Advantages	179
Derating	180
What Can Go Wrong	180
High Temperature Environment	180
Fire Risk	181
Current Inrush	181
Replacement Problems	181
19. neon bulb	183
What It Does	183
How It Works	184
Construction	184
Ionization	184
Negative Resistance	185
How to Use It	186
Limited Light Output	187
Efficiency	187
Ruggedness	187
Power-Supply Testing	188
Life Expectancy	188
Variants	189
Nixie Tubes	189
What Can Go Wrong	189
False Indication	189
Failure in a Dark Environment	189
Premature Failure with DC	190
Premature Failure through Voltage Fluctuations	190
Replacement	190
20. fluorescent light	191
What It Does	191
How It Works	191
Ballast and Starter	192
Flicker	193
Variants	193
CCFLs	194
Sizes	194
Comparisons	194
Values	195
Brightness	195
Spectrum	195
What Can Go Wrong	195
Unreliable Starting	195
Terminal Flicker	195

Cannot Dim	196
Burned Out Electrodes	196
Ultraviolet Hazard	196
21. laser	197
What It Does	197
How It Works	198
Laser Diode	198
Coherent Light	200
Variants	201
CO2 Lasers	201
Fiber Lasers	201
Crystal Lasers	201
Values	201
How to Use It	202
Common Applications	202
What Can Go Wrong	202
Risk of Injury	202
Inadequate Heat Sink	202
Uncontrolled Power Supply	203
Polarity	203
22. LED indicator	205
What It Does	205
Schematic Symbols	206
Common Usage	206
How It Works	207
Multicolor LEDs and Color Mixing	207
Variants	208
Size and Shape	208
Intensity	208
Efficacy	208
Diffusion	209
Wavelength and Color Temperature	209
Internal Resistor	210
Multicolored	210
Infrared	211
Ultraviolet	211
Values	211
Forward Current	211
Low-Current LEDs	211
Forward Voltage	212
Color Rendering Index	212
Life Expectancy	212
Light Output and Heat	212
View Angle	213
How to Use It	213
Polarity	213
Series Resistor Value	214

LEDs in Parallel	214
Multiple Series LEDs	214
Comparisons with Other Light Emitters	214
Other Applications	215
What Can Go Wrong	215
Excessive Forward Voltage	215
Excessive Current and Heat	215
Storage Issues	215
Polarity	215
Internal Resistors	215
23. LED area lighting	217
What It Does	217
Trends in Cost and Efficiency	218
Schematic Symbol	218
How It Works	218
Visible Differences	220
Side-by-Side Comparison	220
Heat Dissipation	222
Efficacy	222
Dimming	222
Ultraviolet Output	222
Color Variation	222
Variants	223
Comparisons	223
Values	225
What Can Go Wrong	225
Wrong Voltage	225
Overheating	225
Fluorescent Ballast Issues	225
Misleading Color Representation	226
 > > MULTI-SOURCE OR PANEL	
24. LED display	227
What It Does	227
How It Works	228
Variants	228
LCD comparisons	228
Seven-Segment Displays	228
Multiple Numerals	229
Additional Segments	229
Dot-Matrix Displays	230
Pixel Arrays	231
Multiple Bar Display	232
Single Light Bar	232
Values	232
How to Use It	232
Seven-Segment Basics	232

Driver Chips and Multiplexing	233
Sixteen-Segment Driver Chip	234
Dot-Matrix LED Display Modules	234
Pixel Arrays	235
Multiple Bar Display Driver	236
One-Digit Hexadecimal Dot Matrix	236
What Can Go Wrong	237
Common Anode versus Common Cathode	237
Incorrect Series Resistance	237
Multiplexing Issues	237
25. vacuum-fluorescent display	239
What It Does	239
How It Works	239
Anode, Cathode, and Grid	240
How to Use It	240
Modern Application	241
Variants	241
Color	241
Character Sets and Pictorial Design	242
Comparisons	242
What Can Go Wrong	242
Fading	242
26. electroluminescence	243
What It Does	243
How It Works	243
Phosphors	244
Derivation	244
Variants	244
Panels	244
Flexible Ribbons	245
Rope Light	245
OLED	246

> SOUND SOURCE

> > AUDIO ALERT

27. transducer	249
What It Does	249
How It Works	249
Variants	250
Electromagnetic	250
Piezoelectric	250
Ultrasonic Transducer	250
Formats	251
Values	251

Frequency Range	251
Sound Pressure	251
Weighted Sound Values	252
Unweighted Values	253
Measurement Location	253
Limitations	253
Voltage	254
Current	254
How to Use It	254
Appropriate Sound Intensity	254
Volume Control	254
AC Supply	254
Self-Drive Transducer Circuit	254
What Can Go Wrong	254
Overvoltage	254
Leakage	255
Component Mounting Problems	255
Moisture	255
Transducer-Indicator Confusion	255
Connection with a Microcontroller	255
28. audio indicator	257
What It Does	257
How It Works	257
Audio Frequency	258
History	258
Variants	258
Sound Patterns	258
Formats	258
Values	259
Voltage	259
Current	260
Frequency	260
Duty Cycle	260
How to Use It	260
Appropriate Sound Intensity	260
Volume Control	260
Wiring	260
What Can Go Wrong	260
> > REPRODUCER	
29. headphone	261
What It Does	261
How It Works	261
Audio Basics	261
Variants	262
Moving Coil	262
Other Types	263

Mechanical Design	264
Values	265
Intensity	265
Frequency Response	265
Distortion	266
Impedance	266
What Can Go Wrong	266
Overdriving	266
Hearing Damage	266
Mismatched Impedance	266
Incorrect Wiring	266
30. speaker	267
What It Does	267
How It Works	267
Construction	268
Multiple Drivers	269
Venting	270
Resonance	270
Miniature Speakers	270
Variants	271
Electrostatic Speaker	271
Powered Speakers	271
Wireless Speakers	271
Innovative Designs	271
Values	271
What Can Go Wrong	272
Damage	272
Magnetic Field	272
Vibration	272
Index	273

How to Use This Book

This is the second of three volumes. Its purpose is to provide an overview of the most commonly used electronic components, for reference by students, engineers, hobbyists, and instructors. While you can find much of this information dispersed among datasheets, introductory books, websites, and technical resources maintained by manufacturers, the *Encyclopedia of Electronic Components* gathers all the relevant facts in one place, properly organized and verified, including details that may be hard to find elsewhere. Each entry includes typical applications, possible substitutions, cross-references to similar devices, sample schematics, and a list of common problems and errors.

You can find a more detailed rationale for this encyclopedia in the Preface to Volume 1.

Volume Contents

Practical considerations influenced the decision to divide this encyclopedia into three volumes. Each deals with broad subject areas as follows.

Volume 1

Power; electromagnetic devices; discrete semiconductors

The *power* category includes sources of electricity and methods to distribute, store, interrupt, convert, and regulate power. The *electromagnet-*

ic devices category includes devices that exert force linearly, and others that create a turning force. *Discrete semiconductors* include the primary types of diodes and transistors. A contents listing for Volume 1 appears in [Figure P-1](#).

Volume 2

Thyristors (SCRs, diacs, and triacs); integrated circuits; light sources, indicators, and displays; and sound sources

Integrated circuits are divided into analog and digital components. *Light sources, indicators, and displays* are divided into reflective displays, single sources of light, and displays that emit light. *Sound sources* are divided into those that create sound, and those that reproduce sound. A contents listing for Volume 2 appears in [Figure P-2](#).

Volume 3

Sensing devices

The field of sensors has become so extensive, they easily merit a volume to themselves. *Sensing devices* include those that detect light, sound, heat, motion, pressure, gas, humidity, orientation, electricity, proximity, force, and radiation.

At the time of writing, Volume 3 is still in preparation, while Volume 1 is complete and is available in a variety of formats.

Primary Category	Secondary Category	Component Type
power	source	battery
	connection	jumper
		fuse
		pushbutton
		switch
		rotary switch
		rotational encoder
	moderation	relay
		resistor
		potentiometer
		capacitor
		variable capacitor
	conversion	inductor
		AC-AC transformer
		AC-DC power supply
		DC-DC converter
		DC-AC inverter
	regulation	voltage regulator
electro-magnetism	linear output	electromagnet
		solenoid
	rotational output	DC motor
		AC motor
		servo motor
		stepper motor
discrete semi-conductor	single junction	diode
		unijunction transistor
	multi-junction	bipolar transistor
		field-effect transistor

Figure P-1. The subject-oriented organization of categories and entries in Volume 1.

Primary Category	Secondary Category	Component Type
discrete semi-conductor	thyristor	SCR
		diac
		triac
integrated circuit	analog	solid-state relay
		optocoupler
		comparator
		op-amp
		digital potentiometer
		timer
	digital	logic gate
		flip-flop
		shift register
		counter
		encoder
		decoder
		multiplexer
light source, indicator or display	reflective	LCD
	single source	incandescent lamp
		neon bulb
		fluorescent light
		laser
		LED indicator
		LED area lighting
	multi-source or panel	LED display
		vacuum-fluorescent
		electroluminescence
sound source	audio alert	transducer
		audio indicator
	reproducer	headphone
		speaker

Figure P-2. The subject-oriented organization of categories and entries in Volume 2.

Organization

Reference versus Tutorial

As its title suggests, this is a reference book, not a tutorial. A tutorial begins with elementary concepts and builds sequentially toward concepts that are more advanced. A reference book assumes that you may dip into the text at any point, learn what you need to know, and then put the book aside. If you choose to read it straight through from beginning to end, you will find some repetition, as each entry is intended to be self-sufficient, requiring minimal reference to other entries.

My books *Make: Electronics* and *Make: More Electronics* follow a tutorial approach. They don't go into as much depth as this Encyclopedia, because a tutorial inevitably allocates a lot of space to step-by-step explanations and instructions.

Theory and Practice

This book is oriented toward practicality rather than theory. I assume that the reader mostly wants to know how to use electronic components, rather than why they work the way they do. Consequently, I have not included proofs of formulae or definitions rooted in electrical theory. Units are defined only to the extent necessary to avoid confusion.

Many books on electronics theory already exist, if theory is of interest to you.

Entries

This encyclopedia is divided into entries, each entry being devoted to one broad type of component. Two rules determine whether a component has an entry all to itself, or is subsumed into another entry:

Rule 1

A component merits its own entry if it is (a) widely used, or (b) not so widely used but has a unique identity and maybe some historical status. The **bipolar transistor** entry is an example of a widely used component,

whereas the **unijunction transistor** entry is an example of a not so widely used component with a unique identity.

Rule 2

A component does not merit its own entry if it is (a) seldom used, or (b) very similar in function to another component that is more widely used. For example, a *rheostat* is subsumed into the **potentiometer** section, while *silicon diode*, *Zener diode*, and *germanium diode* are combined together in the **diode** entry.

Inevitably, these guidelines required judgment calls which in some cases may seem arbitrary. My ultimate decision was based on where I would expect to find a component if I was looking for it myself.

Subject Paths

Entries are not organized alphabetically. They are grouped by subject, in much the same way that books in the nonfiction section of some libraries are organized by the Dewey Decimal System. This is convenient if you don't know exactly what you are looking for, or if you don't know all the options that may be available to perform a task that you have in mind.

Each primary category is divided into subcategories, and the subcategories are divided into component types. This hierarchy is shown in [Figure P-2](#). It is also apparent when you look at the top of the first page of each entry, where you will find the path that leads to it. The **diac** entry, for instance, is headed with this path:

discrete semiconductor > thyristor > diac

Any classification scheme will run into exceptions. You can buy a chip containing a *resistor array*, for instance. Technically, this is an *analog integrated circuit*, but a decision was made to put it in the **resistor** section of Volume 1, because it can be directly substituted for a group of resistors.

Some components have hybrid functions. A **multiplexer**, for instance, may pass analog signals and may have “analog” in its name. However, it is digitally controlled and is mostly used in conjunction with other digital integrated circuits. This seemed to justify placing it in the digital category.

Inclusions and Exclusions

There is also the question of what is, and is not, a component. Is wire a component? Not for the purposes of this encyclopedia. How about a **DC-DC converter**? Because converters are now sold in small packages by component suppliers, they are included in Volume 1 as components.

Many similar decisions had to be made on a case-by-case basis. Some readers will disagree with the outcome, but reconciling all the disagreements would have been impossible. The best I could do was to create a book which is organized in the way that would suit me best if I were using it myself.

Typographical Conventions

Within each entry, **bold type** is used for the first occurrence of the name of a component that has its own entry elsewhere. Other important electronics terms or component names may be presented in *italics*.

The names of components, and the categories to which they belong, are all set in lowercase type, except where a term is normally capitalized because it is an acronym or a trademark. The term *Trimpot*, for instance, is trademarked by Bourns, but *trimmer* is not. **LED** is an acronym, but *cap* (abbreviation for **capacitor**) is not.

The European convention for representing fractional component values eliminates decimal points. Thus, values such as 3.3K and 4.7K are expressed as 3K3 and 4K7. This style has not been adopted to a significant degree in the United States, and is not used in this encyclopedia.

In mathematical formulae, I have used the style that is common in programming languages. The

* (asterisk) is used as a multiplication symbol, while the / (forward slash) is used as a division symbol. Where some terms are in parentheses, they must be dealt with first. Where parentheses are inside parentheses, the innermost ones must be dealt with first. So, in this example:

$$A = 30 / (7 + (4 * 2))$$

You would begin by multiplying 4 times 2, to get 8; then add 7, to get 15; then divide that into 30, to get the value for A, which is 2.

Visual Conventions

Figure P-3 shows the conventions that are used in the schematics in this book. A black dot always indicates a connection, except that to minimize ambiguity, the configuration at top right is avoided, and the configuration at top center is used instead. Conductors that cross each other without a black dot do not make a connection. The styles at bottom right are sometimes seen elsewhere, but are not used here.

All the schematics are formatted with pale blue backgrounds. This enables components such as switches, transistors, and LEDs to be highlighted in white, drawing attention to them and clarifying the boundary of the component. The white areas have no other meaning.

Photographic Backgrounds

All photographs of components include a background grid that is divided into squares measuring 0.1". Although the grid is virtual, it is equivalent in scale to physical graph paper placed immediately behind the component. If the component is photographed at an angle, the grid may be reproduced at a similar angle, creating perspective on the squares.

Background colors in photographs were chosen for contrast with the colors of the components, or for visual variety. They have no other significance.

Component Availability

Because there is no way of knowing if a component may have a long production run, this encyclopedia is cautious about listing specific part numbers. To find a specific part that has a narrow function, searching the websites maintained by suppliers will be necessary. The following suppliers were checked frequently during the preparation of the book:

- [Mouser Electronics](#)
- [Jameco Electronics](#)

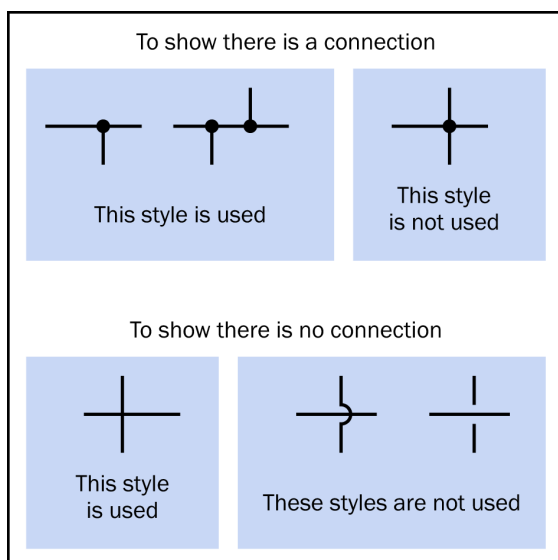


Figure P-3. Visual conventions that are used in the schematics in this book.

When seeking obsolete parts, or those that are nearing the end of their commercial life, eBay can be very useful.

Issues and Errata

If you believe you have found an error in this book, you will find guidance on how to report it here: http://bit.ly/eec_v2_errata.

Before posting your own erratum, please check those that have been submitted previously, to see if someone else already reported it.

I value and encourage reader feedback. However, before you post feedback publicly to a site such as Amazon, I have a request. Please be aware of the power that you have as a reader, and use it fairly. A single negative review can create a bigger effect than you may realize. It can certainly outweigh half-a-dozen positive reviews. If you feel you have not received a prompt or adequate response from the O'Reilly errata site mentioned here, you can email me personally at:

make.electronics@gmail.com

I check that address irregularly—sometimes only once in a couple of weeks. But I do answer all messages.

Safari® Books Online

Safari Books Online is an on-demand digital library that delivers expert [content](#) in both book and video form from the world's leading authors in technology and business.

Technology professionals, software developers, web designers, and business and creative professionals use Safari Books Online as their primary resource for research, problem solving, learning, and certification training.

Safari Books Online offers a range of [plans and pricing](#) for [enterprise](#), [government](#), [education](#), and individuals.

Members have access to thousands of books, training videos, and prepublication manuscripts in one fully searchable database from publishers like Maker Media, O'Reilly Media, Prentice Hall Professional, Addison-Wesley Professional, Microsoft Press, Sams, Que, Peachpit Press, Focal Press, Cisco Press, John Wiley & Sons, Syngress, Morgan Kaufmann, IBM Redbooks, Packt, Adobe Press, FT Press, Apress, Manning, New Riders, McGraw-Hill, Jones & Bartlett, Course Technology, and hundreds [more](#). For more information about Safari Books Online, please visit us [online](#).

How to Contact Us

Please address comments and questions concerning this book to the publisher:

Make:
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (in the United States or Canada)
707-829-0515 (international or local)
707-829-0104 (fax)

Make: unites, inspires, informs, and entertains a growing community of resourceful people who undertake amazing projects in their backyards, basements, and garages. Make: celebrates your right to tweak, hack, and bend any technology to your will. The Make: audience continues to be a growing culture and community that believes in bettering ourselves, our environment, our educational system—our entire world. This is much more than an audience, it's a worldwide movement that Make: is leading—we call it the Maker Movement.

For more information about Make:, visit us online:

Make: magazine: <http://makezine.com/magazine/>
Maker Faire: <http://makerfaire.com>
Makezine.com: <http://makezine.com>
Maker Shed: <http://makershed.com/>

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at: http://bit.ly/encyclopedia_of_electronic_components_v2.

Acknowledgments

Any reference work draws inspiration from many sources. Datasheets and tutorials maintained by component manufacturers were considered the most trustworthy sources of information online. In addition, component retailers, college texts, crowd-sourced reference works, and hobbyist

sites were used. The following books provided useful information:

Boylestad, Robert L. and Nashelsky, Louis: *Electronic Devices and Circuit Theory*, 9th edition. Pearson Education, 2006.

Braga, Newton C.: *CMOS Sourcebook*. Sams Technical Publishing, 2001.

Hoenig, Stuart A.: *How to Build and Use Electronic Devices Without Frustration, Panic, Mountains of Money, or an Engineering Degree*, 2nd edition. Little, Brown, 1980.

Horn, Delton T.: *Electronic Components*. Tab Books, 1992.

Horn, Delton T.: *Electronics Theory*, 4th edition. Tab Books, 1994.

Horowitz, Paul and Hill, Winfield: *The Art of Electronics*, 2nd edition. Cambridge University Press, 1989.

Ibrahim, Dogan: *Using LEDs, LCDs, and GLCDs in Microcontroller Projects*. John Wiley & Sons, 2012.

Kumar, A. Anand: *Fundamentals of Digital Circuits*, 2nd edition. PHI Learning, 2009.

Lancaster, Don: *TTL Cookbook*. Howard W. Sams & Co, 1974.

Lenk, Ron and Lenk, Carol: *Practical Lighting Design with LEDs*. John Wiley & Sons, 2011.

Lowe, Doug: *Electronics All-in-One for Dummies*. John Wiley & Sons, 2012.

Mims III, Forrest M.: *Getting Started in Electronics*. Master Publishing, 2000.

Mims III, Forrest M.: *Electronic Sensor Circuits & Projects*. Master Publishing, 2007.

Mims III, Forrest M.: *Timer, OpAmp, & Optoelectronic Circuits and Projects*. Master Publishing, 2007.

Predko, Mike: *123 Robotics Experiments for the Evil Genius*. McGraw-Hill, 2004.

Scherz, Paul: *Practical Electronics for Inventors*, 2nd edition. McGraw-Hill, 2007.

Williams, Tim: *The Circuit Designer's Companion*, 2nd edition. Newnes, 2005.

I also made extensive use of information on vendor sites, especially:

- [Mouser Electronics](#)
- [Jameco Electronics](#)
- [All Electronics](#)
- [sparkfun](#)
- [Electronic Goldmine](#)
- [Adafruit](#)
- [Parallax, Inc.](#)

In addition, some individuals provided special assistance. My editor, Brian Jepson, was im-

mensely helpful in the development of this book. Philipp Marek and Steve Conklin reviewed the text for errors. My publisher demonstrated faith in my work. Kevin Kelly unwittingly influenced me with his legendary interest in “access to tools.” It was Mark Frauenfelder who originally brought me back to the pleasures of building things, and Gareth Branwyn who revived my interest in electronics.

Lastly, I should mention my school friends from decades ago: Patrick Fagg, Hugh Levinson, Graham Rogers, William Edmondson, and John Witty, who helped me to feel that it was OK to be a nerd building my own audio equipment, long before the word “nerd” actually existed.

—Charles Platt, 2014

SCR

1

The acronym **SCR** is derived from *silicon-controlled rectifier*, which is a gate-triggered type of *thyristor*. A thyristor is defined here as a semiconductor having four or more alternating layers of p-type and n-type silicon. Because it predated integrated circuits, and in its basic form consists of a single multilayer semiconductor, a thyristor is considered to be a discrete component in this encyclopedia. When a thyristor is combined with other components in one package (as in a **solid-state relay**), it is considered to be an integrated circuit.

Other types of thyristor are the **diac** and **triac**, each of which has its own entry.

Thyristor variants that are not so widely used, such as the *gate turn-off thyristor (GTO)* and *silicon-controlled switch (SCS)*, do not have entries here.

OTHER RELATED COMPONENTS

- **diac** (see [Chapter 2](#))
- **triac** (see [Chapter 3](#))

What It Does

In the 1920s, the *thyatron* was a gas-filled tube that functioned as a switch and a rectifier. In 1956, General Electric introduced a solid-state version of it under the name *thyristor*. In both cases, the names were derived from the thyroid gland in the human body, which controls the rate of consumption of energy. The thyatron and, subsequently, the thyristor enabled control of large flows of current.

The **SCR** (silicon-controlled rectifier) is a type of thyristor, although the two terms are often used as if they are synonymous. Text that refers loosely to a thyristor may actually be discussing an SCR, and vice versa. In this encyclopedia, the **SCR**, **diac**, and **triac** are all considered to be variant types of thyristor.

An SCR is a solid-state switch that in many instances can pass high currents at high voltages.

Like a **bipolar transistor**, it is triggered by voltage applied to a gate. Unlike the transistor, it allows the flow of current to continue even when the gate voltage diminishes to zero.

How It Works

This component is designed to pass current in one direction only. It can be forced to conduct in the opposite direction if the reversed potential exceeds its *breakdown voltage*, but this mistreatment is likely to cause damage.

By comparison, the diac and triac are designed to be bidirectional.

The SCR has three leads, identified as anode, cathode, and gate. Two functionally identical versions of the schematic symbol are shown in [Figure 1-1](#). Early versions sometimes included a circle drawn around them, but this style has become obsolete. Care must be taken to distinguish

between the SCR symbol and the symbol that represents a **programmable unijunction transistor** (PUT), shown in Figure 1-2.

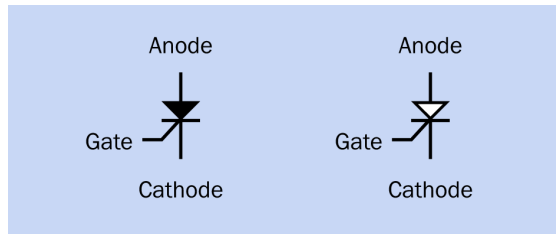


Figure 1-1. Two functionally identical schematic symbols for an SCR (silicon-controlled rectifier). The symbol on the left is more common.

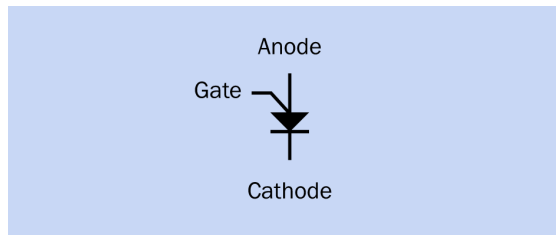


Figure 1-2. The symbol shown here is for a programmable unijunction transistor (PUT). Care must be taken to distinguish it from the symbol for an SCR.

Switching Behavior

When the SCR is in its passive or nonconductive state, it will block current in either direction between anode and cathode, although a very small amount of *leakage* typically occurs. When the SCR is activated by a positive voltage at the gate, current can now flow from anode to cathode, although it is still blocked from cathode to anode. When the flow reaches a level known as the *latching current*, the flow will continue even after the triggering voltage drops to zero. This behavior causes it to be known as a *regenerative* device.

If the current between anode and cathode starts to diminish while the gate voltage remains zero, the current flow will continue below the latching level until it falls below the value known as the *holding current*. The flow now ceases. Thus, the only way to end a flow of current that has been

initiated through an SCR is by reducing the flow or attempting to reverse it.

Note that the self-sustaining flow is a function of current rather than voltage.

Unlike a transistor, an SCR is either “on” or “off” and does not function as a *current amplifier*. Like a diode, it is designed to conduct current in one direction; hence the term *rectifier* in its full name. When it has been triggered, the impedance between its anode and cathode is sufficiently low that heat dissipation can be managed even at high power levels.

The ability of SCRs to pass relatively large amounts of current makes them suitable for controlling the power supplied to motors and resistive heating elements. The fast switching response also enables an SCR to interrupt and abbreviate each positive phase of an AC waveform, to reduce the average power supplied. This is known as *phase control*.

SCRs are also used to provide *overvoltage protection*.

SCR packages reflect their design for a wide range of voltages and currents. Figure 1-3 shows an SCR designed for on-state current of 4A RMS (i.e., measured as the root mean square of the alternating current). Among its applications are small-engine ignition and *crowbar* overvoltage protection, so named because it shorts a power supply directly to ground, much like a crowbar being dropped across the terminals of a car battery (but hopefully with a less dramatic outcome). See Figure 1-15.

In Figure 1-4, the SCR can handle up to 800V repetitive peak off-state voltage and 55A RMS. Possible applications include AC rectification, crowbar protection, welding, and battery charging. The component in Figure 1-5 is rated for 25A and 50V repetitive peak off-state voltage. To assess the component sizes, bear in mind that the graph line spacing is 0.1”.

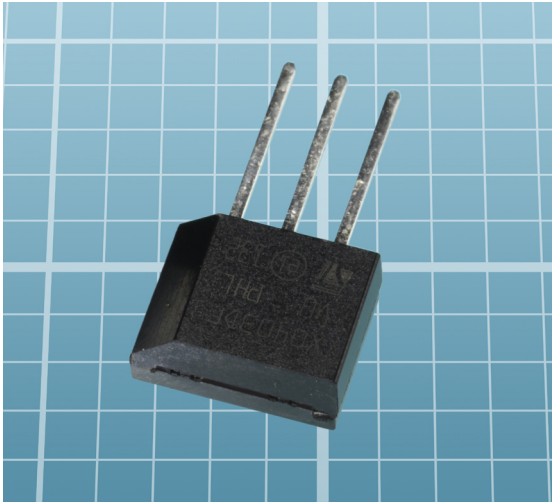


Figure 1-3. SCR rated for 400V repetitive off-state voltage, no greater than 4A RMS.

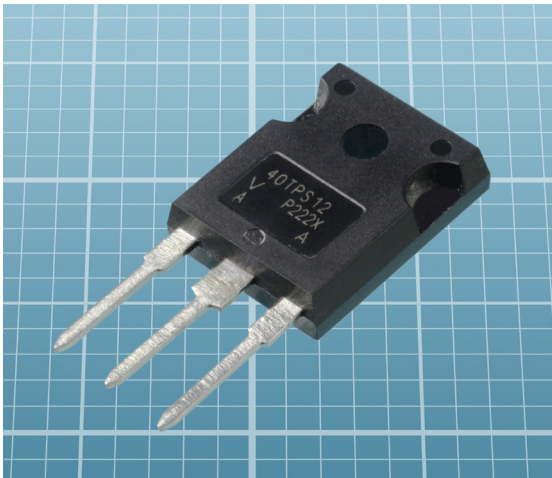


Figure 1-4. SCR rated for 800V repetitive off-state voltage, no greater than 55A RMS.

Internal Configuration

The function of an SCR can be imagined as being similar to that of a PNP transistor paired with an NPN transistor, as shown in [Figure 1-6](#). In this simplified schematic, so long as zero voltage is applied to the “gate” wire, the lower (NPN) transistor remains nonconductive. Consequently, the upper (PNP) transistor cannot sink current, and this transistor also remains nonconductive.

When voltage is applied to the “gate,” the lower transistor starts to sink current from the upper transistor. This switches it on. The two transistors now continue to conduct even if power to the “gate” is disconnected, because they have created a positive feedback loop.



Figure 1-5. Stud-packaged SCR rated for 50V repetitive off-state voltage, no greater than 25A RMS.

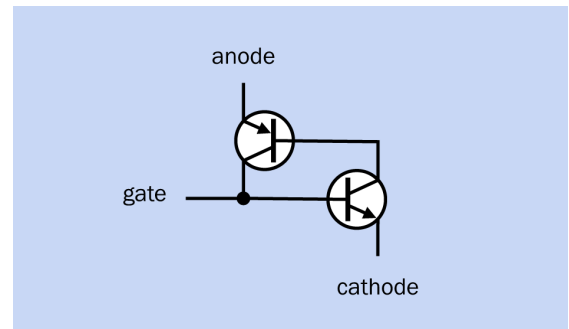


Figure 1-6. An SCR behaves similarly to an NPN and a PNP transistor coupled together.

[Figure 1-7](#) shows the same two transistors in simplified form as sandwiches of p-type and n-type silicon layers (on the left), and their combination in an SCR (on the right). Although the actual configuration of silicon segments is not as simple or as linear as this diagram suggests, the SCR can be described correctly as a [PNPN device](#).

An SCR is comparable with an electromagnetic **latching relay**, except that it works much faster and more reliably.

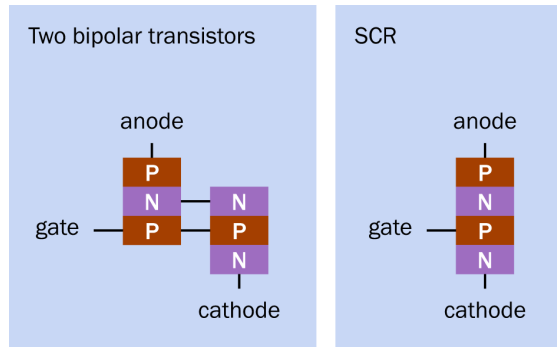


Figure 1-7. The two transistors from the previous figure are shown here in simplified form as two stacks of p-type and n-type silicon layers. These layers are combined in an SCR, on the right.

Breakdown and Breakover Voltage

The curves in [Figure 1-8](#) illustrate the behavior of a hypothetical SCR, and can be compared with the curves shown for a diac in [Figure 2-5](#) and a triac in [Figure 3-10](#). Beginning with zero voltage applied between anode and cathode, and zero current flowing (i.e., at the center origin of the graph), if we apply a voltage at the anode that is increasingly negative relative to the cathode (i.e., we attempt to force the SCR to allow negative current flow), we see a small amount of leakage, indicated by the darker blue area (which is not drawn to scale). Finally the **breakdown voltage** is reached, at which point the negative potential overcomes the SCR and its impedance drops rapidly, allowing a surge of current to flow, probably damaging it.

Alternatively, starting once again from the center, if we apply a voltage at the anode that is increasingly positive relative to the cathode, two consequences are possible. The dashed curve assumes that there is zero voltage at the gate, and shows that some leakage occurs until the applied potential at the anode reaches the **breakover voltage**, at which point the SCR allows a large

current flow, which continues even when the voltage decreases.

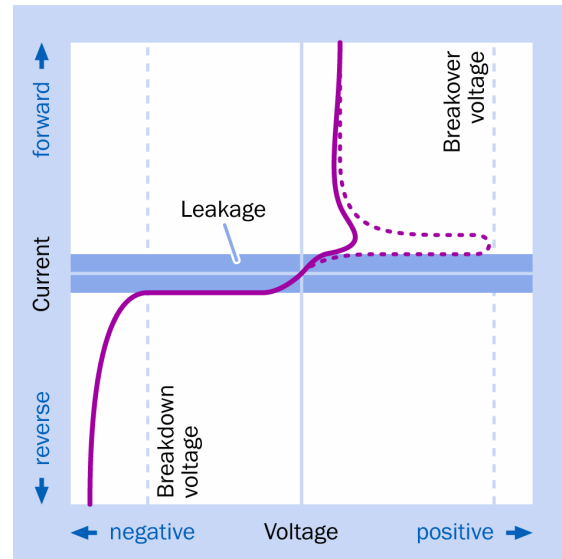


Figure 1-8. The solid curve shows current passing between the anode and cathode of a hypothetical SCR for varying voltages, while a triggering voltage is applied to the gate. The dashed curve assumes that no triggering voltage is applied to the gate.

In practice, the SCR is intended to respond to a positive gate voltage. Under these circumstances, its behavior is shown by the solid curve in the top-right quadrant in [Figure 1-8](#). The SCR begins to conduct current without having to reach the breakover voltage at the anode.

- When used as it is intended, the SCR should not reach breakdown or breakover voltage levels.

SCR Concept Demo

In [Figure 1-9](#), pushbutton S1 applies voltage to the gate of the SCR, which puts the SCR in self-sustaining conductive mode. When S1 is released, the meter will show that current continues to pass between the anode and the cathode. The X0403DF SCR suggested for this circuit has a holding current of 5mA, which a 5VDC supply should be able to provide with the 1K resistor in

the circuit. If necessary, this resistor can be reduced to 680Ω .

Now if pushbutton S2 is pressed, the flow is interrupted. When S2 is released, the flow will not resume. Alternatively, if pushbutton S3 is pressed while the SCR is conducting current, the flow is diverted around the SCR, and when the pushbutton is released, the flow through the SCR will not resume. Thus, the SCR can be shut down either by a normally closed pushbutton in series with it (which will interrupt the current), or a normally open pushbutton in parallel with it (which will divert the current).

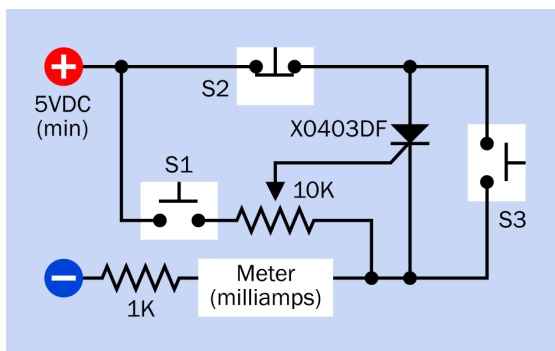


Figure 1-9. In this test circuit, S1 triggers the SCR, while S2 or S3 will stop it. See text for additional details.

The test circuit is shown installed on a breadboard in [Figure 1-10](#). In this photograph, the red and blue wires supply a minimum of 5VDC. The two red buttons are tactile switches, the one at top left being S1 in the schematic while the one at bottom right is S3. The large switch with a rectangular button is S2; this is normally closed, and opens when pressed. The X0403DF SCR is just below it and to the right. The square blue trimmer is set to the midpoint of its range.

AC Current Applications

If the SCR is used with alternating current, it stops conducting during each negative cycle, and is retriggered in each positive cycle. This suggests one of its primary applications, as a controllable rectifier that can switch rapidly enough to limit

the amount of current that passes through it during each cycle.

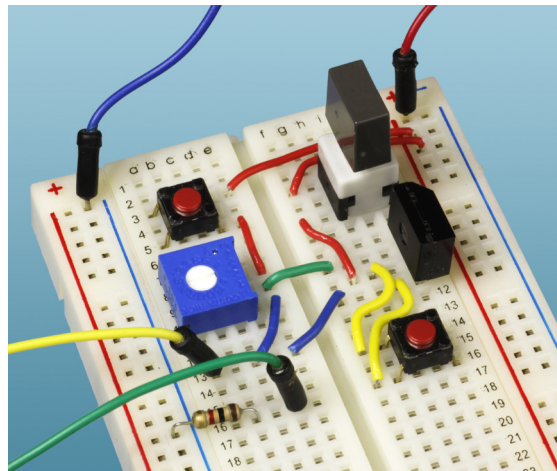


Figure 1-10. A breadboarded version of the SCR test circuit. The two red buttons correspond with S1 and S3 in the schematic, while the large rectangular button at top right opens S2. See text for details.

Variants

SCRs are available in surface-mount, through-hole, and stud packages, to handle increasing currents and voltages. Some special-purpose SCRs can control currents of hundreds of amps, while high-power SCRs are used to switch thousands of amps at more than 10,000V in power distribution systems. They are too specialized for inclusion in this encyclopedia.

Typical power ratings for SCRs in general use are summarized in the next section.

Values

Any SCR will impose a forward voltage drop, which typically ranges from around 1V to 2V, depending on the component.

Because SCRs are often used to modify AC waveforms, the current that the component can pass is usually expressed as the root mean square (RMS) of its peak value.

Commonly Used Abbreviations

- V_{DRM} Maximum repetitive forward voltage that may be applied to the anode while no voltage is applied to the gate (i.e., when the SCR is not in conductive mode).
- V_{RRM} Maximum repetitive reverse voltage that may be applied to the anode while no voltage is applied to the gate (i.e., when the SCR is not in conductive mode).
- V_{TM} Maximum on-state voltage while the SCR is in conductive mode. T indicates that this value changes with temperature.
- V_{GM} Forward maximum gate voltage.
- V_{GT} Minimum gate voltage required to trigger.
- V_{GD} Maximum gate voltage that will not trigger.
- I_{DRM} Peak repetitive forward blocking current (i.e., maximum leakage).
- I_{RRM} Peak repetitive reverse blocking current (i.e., leakage in the off state).
- I_{GM} Maximum forward gate current.
- $I_{\text{T(RMS)}}$ Maximum RMS current between anode and cathode while the SCR is in conductive mode. T indicates that this value changes with temperature.
- $I_{\text{T(AV)}}$ Maximum average current between anode and cathode while the SCR is in conductive mode. T indicates that this value changes with temperature.
- I_{GT} Maximum gate current required to trigger.
- I_{H} Typical holding current.
- I_{L} Maximum latching current.
- T_{C} Case temperature, usually expressed as an acceptable range.
- T_{J} Operating junction temperature, usually expressed as an acceptable range.

Surface-mount variants may tolerate maximum anode-cathode currents that typically range from 1A to 10A. Maximum voltages as high as 500V are allowed in some cases. Leakage in the “off” state may be as high as 0.5mA or as low as 5 μ A. Gate trigger voltage is likely to range from 0.8V to 1.5V, and trigger current of 0.2mA to 15mA is typical.

Through-hole variants may be packaged in TO-92 format (like discrete transistors) or, more commonly, in TO-220 format (like a typical 1A voltage regulator). They may be rated for a maximum of 5A up to 50A, depending on the component, with maximum voltages ranging from 50V to 500V. Leakage is similar to surface-mount variants. The gate trigger voltage is typically around 1.5V, and trigger current ranges from 25–50mA.

A stud-type SCR may have a maximum 50A to 500A current rating, although some components are capable of tolerating even higher values. Maximum voltages of 50V to 500V are possible. Leakage is likely to be higher than in other formats, with 5mA to 30mA being common. The gate trigger voltage is typically 1.5 to 3V, and trigger current may range from around 50mA to 200mA.

How to Use It

Although other applications are possible, in practice SCRs have two main applications:

- Phase control, which interrupts each positive phase of an AC power supply. It can moderate the speed of a motor or the heat generated by a resistive load.
- Overvoltage protection. This can safeguard sensitive components in a circuit where there is a DC power supply.

SCRs are often incorporated in ground-fault circuit interruptors (although not usually as discrete components) and in automotive ignition systems.

Phase Control

Phase control is a convenient way to control or limit the AC power delivered to a load by abbreviating each pulse in the AC waveform. This is done by adjusting the gate voltage so that the SCR blocks the first part of each positive phase, then conducts the remainder, and then stops conducting below its holding level. The SCR will then block the reversed flow in the negative phase of the AC waveform, but an additional SCR with opposite polarity can be added.

This is a form of *pulse-width modulation*. It is highly efficient, as the effective internal resistance of the SCR is either very high or very low, and the component does not waste significant energy in the form of heat.

On a graph showing the fluctuating voltage of an AC waveform, a single cycle is customarily divided into four stages: (1) zero voltage, (2) maximum positive voltage, (3) zero voltage, (4) minimum negative voltage, all measurements being made between the live side of the supply and the neutral side of the supply.

The cycle then repeats. Its transitions are often referred to as *phase angles* of 0 degrees, 90 degrees, 180 degrees, and 270 degrees, as shown in [Figure 1-11](#).

The fluctuating voltage in an AC power supply is proportional with the sine of the phase angle. This concept is illustrated in [Figure 1-12](#). If an imaginary point (shown as a purple dot) is moving in a circular path counterclockwise at a constant speed, its vertical distance (shown in green) above or below the X axis (horizontal centerline) can represent an AC voltage corresponding with the angle (shown as purple arcs) of the circle radius to the point, each angle being measured from at the center relative to a start position at right on the X axis.

When an SCR is used for phase control, the point at which it starts to conduct may be anywhere from 0 to almost 180 degrees. This is achieved by diverting a small amount of the AC power into an RC network attached to the gate of the SCR, as

shown in [Figure 1-13](#). The capacitor in this schematic introduces a delay that can be varied by the potentiometer. This enables the SCR to be triggered even after the peak of the AC power signal. In [Figure 1-14](#), the AC power is shown by the center (green) curve, and the slightly delayed, reduced voltage at the gate is shown by the upper, purple curve. When the gate voltage rises to the trigger level, it causes the SCR to begin conducting current, creating an abbreviated output shown in the bottom curve. In this way, triggering from an AC phase angle of 0 degrees to almost 180 degrees is possible. The phase angle where the SCR begins to allow conduction is known as the *conduction angle*.

If two SCRs with opposite polarity are placed in parallel with each other, they can be used to provide phase control on both the positive-going and negative-going sections of an AC cycle. This configuration is used in high-powered devices. A **triac** is used for the same purpose with lower current.

Six SCRs may be used to control three-phase power.

Overvoltage Protection

The tolerance of an SCR for high current makes it suitable for use in a crowbar voltage limiting circuit.

In [Figure 1-15](#), the SCR does not conduct current (other than a small amount of leakage) until the Zener diode senses a voltage above the maximum level considered safe. The diode then allows power to reach the gate of the SCR. Its impedance drops immediately, and the resulting surge of current trips the fuse. After the cause of the overvoltage condition is corrected, the fuse can be replaced and the circuit may resume functioning.

A capacitor is included so brief spikes in the power supply will be passed to ground without triggering the SCR. A resistor of around 100Ω ensures that the gate voltage of the SCR remains near zero during normal operation. When the Zener

diode starts to conduct current, the resistor acts as a voltage divider with the diode, so that sufficient voltage reaches the SCR to activate it.

This circuit may be unsuitable for low-voltage power supplies, because the Zener diode has to be chosen with a high enough rating to prevent small power fluctuations from tripping it. Bearing in mind that the real triggering voltage of the diode may be at least plus-or-minus 5% of its rated voltage, the diode may have to be chosen with at least a 6V rating in a 5V circuit, and it may not be activated until the voltage is actually 6.5V. This may be insufficient to protect the components being used with the power supply.

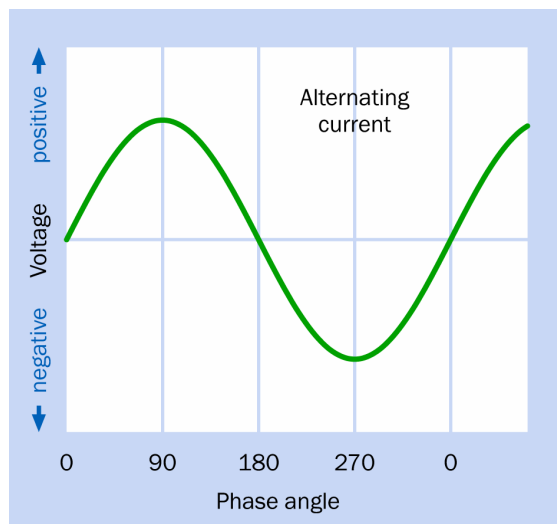


Figure 1-11. An AC waveform is customarily measured in degrees of phase angle.

What Can Go Wrong

Like other semiconductors, an SCR can be adversely affected by excessive heat. Usual precautions should be taken to allow sufficient ventilation and heat sinking, especially when components are moved from an open prototyping board to an enclosure in which crowding is likely.

Unexpected Triggering Caused by Heat

On a datasheet, the values for triggering current and holding current are valid only within a recommended temperature range. A buildup of heat can provoke unexpected triggering.

Unexpected Triggering Caused by Voltage

A very rapid increase in forward voltage at the anode can induce a triggering voltage in the gate by capacitive coupling. As a result, the SCR can trigger itself without any external application of gate voltage. This is sometimes known as *dv/dt triggering*. If necessary, a snubber circuit can be added across the anode input to prevent sudden voltage transitions.

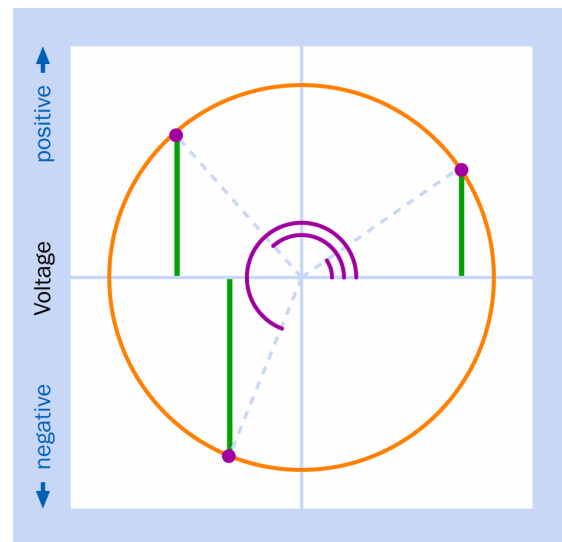


Figure 1-12. The fluctuating voltage of an AC power supply (shown as vertical green lines) is proportional with the sines of the angles (purple arcs) in this diagram. The angles are referred to as phase angles.

Confusion of AC and DC Ratings

The on-state current for an SCR is averaged only over the width of each pulse that the SCR actually conducts. It is not time-averaged over an entire AC cycle, and it will be different again from a DC rating. Care must be taken to match the current

rating with the way in which the component will actually be used.

Maximum Current versus Conduction Angle

Current-carrying capability will be very significantly affected by the length of the duty cycle when the SCR is being used to abbreviate each positive AC pulse. When the SCR imposes a 120-degree conduction angle, it may be able to handle twice the average on-state current as when it is imposing a 30-degree conduction angle. The manufacturer's datasheet should include a graphical illustration of this relationship. If an SCR is chosen for a high conduction angle, and the angle is later reduced, overheating will result, and damage is likely.

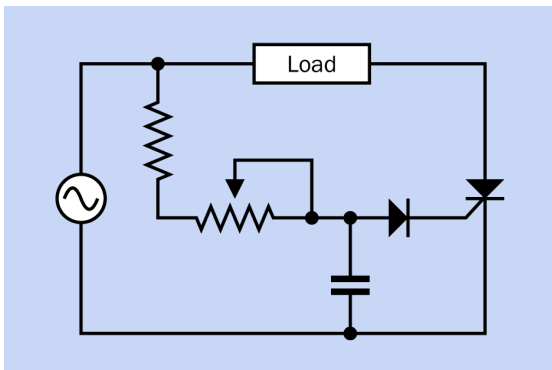


Figure 1-13. In this schematic, an SCR is used to apply phase control, adjusting the power that passes through a load.

Confusing Symbols

When reading a schematic, unfortunate errors can result from failure to distinguish between the symbol for a **programmable unijunction transistor (PUT)** and the symbol for an SCR. The characteristics of a PUT are described in Volume 1 of this encyclopedia.

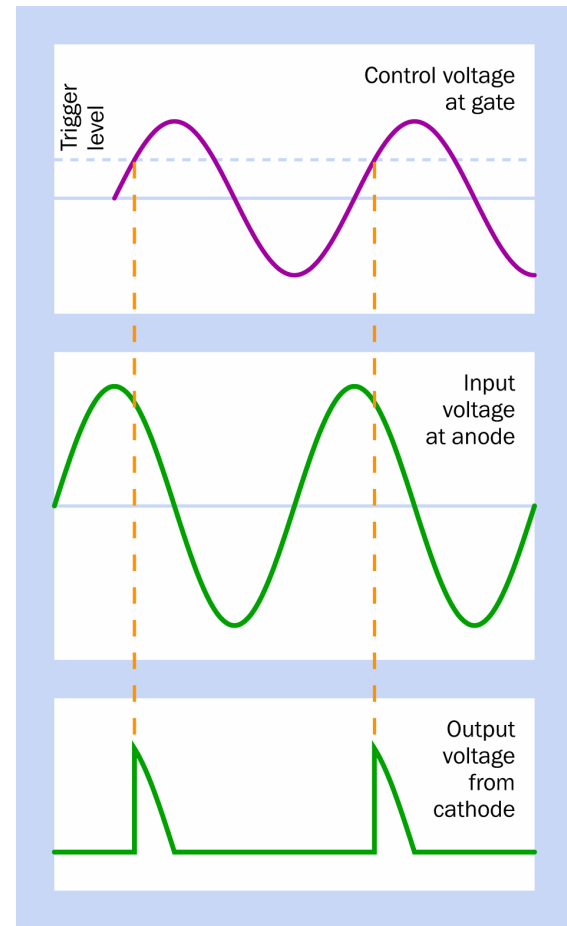


Figure 1-14. If the AC power applied to the anode of an SCR (center) is reduced in voltage and delayed slightly by an RC network, it can trigger the SCR (top), causing it to pass only an abbreviated segment of each positive AC pulse (bottom).

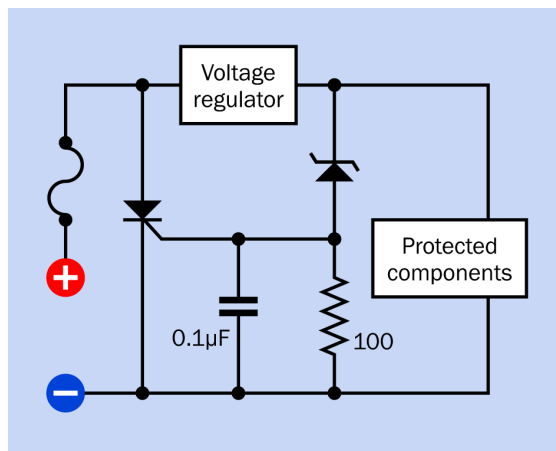


Figure 1-15. In this schematic, an SCR is used to provide crowbar overvoltage protection for sensitive components.

diac

2

A **diac** is a self-triggering type of *thyristor*. Its name is said to be derived from the phrase “diode for AC,” and because it is not an acronym, it is not usually capitalized.

A thyristor is defined here as a semiconductor having four or more layers of p-type and n-type silicon. Because the thyristor predated integrated circuits, and in its basic form consists of a single multilayer semiconductor, it is categorized as a discrete component in this encyclopedia. When a thyristor is combined with other components in one package (as in a **solid-state relay**) it is considered to be an integrated circuit.

Other types of thyristor are the **SCR** (silicon-controlled rectifier) and the **triac**, each of which has its own entry in this encyclopedia.

Thyristor variants that are not so widely used, such as the *gate turn-off thyristor (GTO)* and *silicon-controlled switch (SCS)*, do not have entries here.

OTHER RELATED COMPONENTS

- **SCR** (see [Chapter 1](#))
- **triac** (see [Chapter 3](#))

What It Does

The **diac** is a bidirectional thyristor with only two terminals. It blocks current until it is subjected to sufficient voltage, at which point its impedance drops very rapidly. It is primarily used to trigger a **triac** for purposes of moderating AC power to an **incandescent lamp**, a resistive heating element, or an AC **motor**. The two leads on a diac have identical function and are interchangeable.

By comparison, a **triac** and an **SCR** are thyristors with three leads, one of them being referred to as the gate, which determines whether the component becomes conductive. A triac and a diac allow current to flow in either direction, while an SCR always blocks current in one direction.

Symbol Variants

The schematic symbol for a diac, shown in [Figure 2-1](#), resembles two diodes joined together, one of them inverted relative to the other. Functionally, the diac is comparable with a pair of Zener diodes, as it is intended to be driven beyond the point where it becomes saturated. Because its two leads are functionally identical, they do not require names to differentiate them. They are sometimes referred to as A1 and A2, in recognition that either of them may function as an anode; or they may be identified as MT1 and MT2, MT being an acronym for “main terminal.”

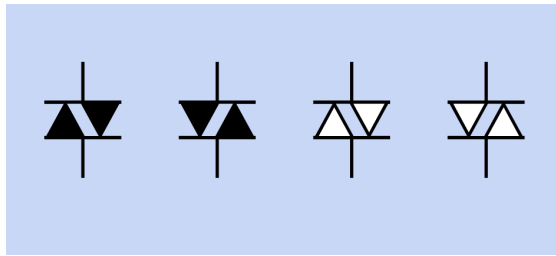


Figure 2-1. Symbol variants to represent a diac. All four are functionally identical.

The symbol may be reflected left to right, and the black triangles may have open centers. All of these variants mean the same thing. Occasionally the symbol has a circle around it, but this style is now rare.

When only a moderate voltage is applied (usually less than 30V) the diac remains in a passive state and will block current in either direction, although a very small amount of *leakage* typically occurs. When the voltage exceeds a threshold known as its *breakover level*, current can flow, and the diac will continue to conduct until the current falls below its *holding level*.

A sample diac is shown in Figure 2-2.

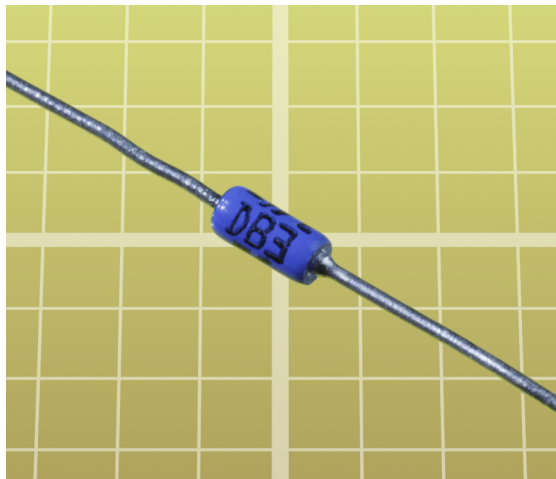


Figure 2-2. Because a diac is not intended to pass significant current, it is typically packaged in a small format. The graph squares in the photograph each measure 0.1".

How It Works

Figure 2-3 shows a circuit that demonstrates the conductive behavior of a diac.

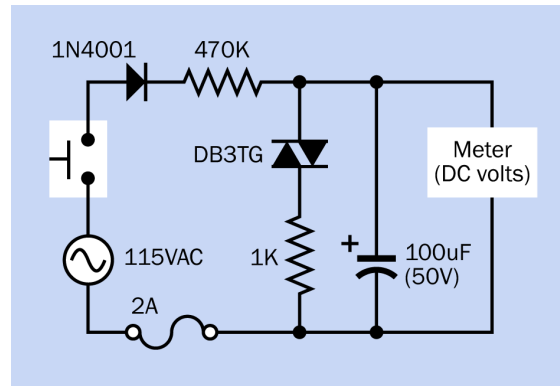


Figure 2-3. A test circuit to demonstrate the behavior of a diac. See text for details.

When the pushbutton is held down, current from the positive side of the AC supply flows through the diode and the 470K resistor to the capacitor. The diac is not yet conductive, so the capacitor accumulates a potential that can be monitored with the volt meter. After about 30 seconds, the charge on the capacitor reaches 32V. This is the breakover voltage for this particular diac, so it becomes conductive. The positive side of the capacitor can now discharge through the diac and the 1K series resistor to ground.

If the pushbutton is released at this moment, the meter will show that the capacitor discharges to a potential below the holding level of the diac. The capacitor now stops discharging because the diac has ceased being conductive.

If the pushbutton is held down constantly, the meter will show the capacitor charging and then discharging through the diac repeatedly, so that the circuit behaves as a *relaxation oscillator*. The 1K series resistor is included to protect the diac from excessive current. If a standard quarter-watt resistor is used, it should not become unduly warm because current passes through it only intermittently.

- Because this circuit uses 115VAC, basic precautions should be taken. The fuse should not be omitted, the capacitor should be rated for at least 50V, and the circuit should not be touched while it is connected to the power source. Breadboarding a circuit using this voltage requires caution and experience, as wires can easily come loose, and components can be touched accidentally while they are live.

Figure 2-4 shows the test circuit on a breadboard. The red and blue leads at the top of the photograph are from a fused 115VAC power supply. The live side of the supply passes through a diode to a pushbutton switch that has a rectangular black cap. A 470K resistor connects the other side of the switch to the positive side of a 100 μ F electrolytic capacitor, and also to the diac (small blue component). A 1K resistor connects the other end of the diac back to the negative side of the capacitor, which is grounded. The yellow and blue wires leaving the photograph at the left are connected with a volt meter, which is not shown.

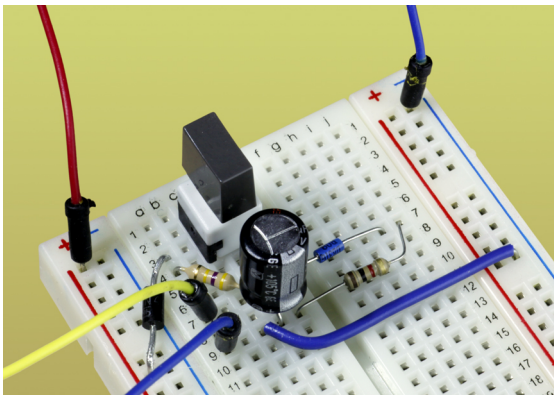


Figure 2-4. A breadboarded version of the diac test circuit. See text for details.

The behavior of a diac is also illustrated in Figure 2-5, which can be compared with the curves in Figures 3-10 and 1-8, depicting the behavior of a triac and an SCR respectively.

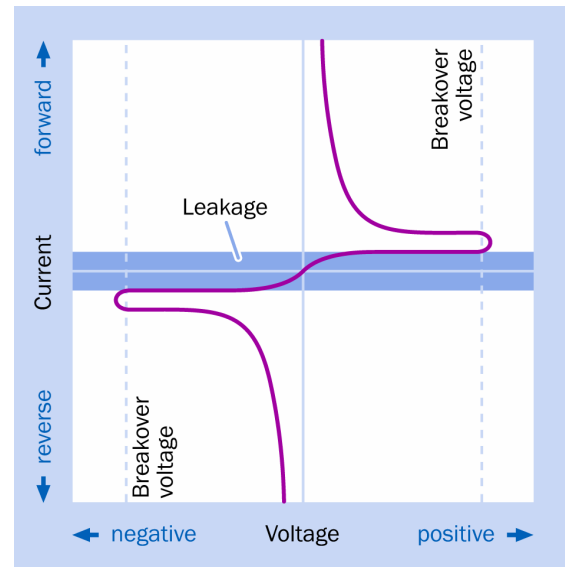


Figure 2-5. The curve shows current passing through a diac when various voltages are applied.

Switching AC

The diac cannot function as a switch, because it lacks the third terminal which is found in a triac, an SCR, or a bipolar transistor. However, it is well suited to drive the gate of a triac, because the behavior of a diac is symmetrical in response to opposite voltages, while the triac is not. If an AC voltage applied to a diac is adjusted with a potentiometer in an RC circuit, the diac will pass along a portion of each positive or negative pulse, and will delay it by a brief amount of time determined by the value of the capacitor in the RC circuit and the setting of the potentiometer. This is known as *phase control*, as it controls the *phase angle* at which the diac allows current to flow.

See Figure 3-13 for a schematic showing a diac driving a triac. See Figures 1-14 and 3-11 for graphs illustrating phase control. See “Phase Control” for a discussion of phase in AC waveforms.

Variants

Diacs are available in through-hole and surface-mount formats. Because they are not intended

to handle significant current, no heat sink is included.

A *sidac* behaves very similarly to a diac, its name being derived from “silicon diode for alternating current.” Its primary difference from generic diacs is that it is designed to reach its breakover voltage at a higher value, typically 120VAC or 240VAC.

Values

When performing its function to trigger a triac, a diac is unlikely to pass more than 100mA.

The breakover voltage of a diac is usually between 30V and 40V, with a few versions designed for up to 70V. When the diac starts to conduct, its on-state impedance is sufficient to reduce the voltage significantly, with 5V being a typical minimum output voltage.

Although the rise time when a diac responds is very brief (around 1μs), the component is not expected to run at a high frequency. It will normally be used with 50Hz or 60Hz AC to trigger a triac. For this reason, its repetitive peak on-state current is usually specified at no more than 120Hz.

Abbreviations in datasheets are likely to include:

- V_{BO} Breakover voltage (sometimes may be specified as latching voltage, which for a diac is the same thing).
- $V_{BO1} - V_{BO2}$ Breakover voltage symmetry. The hyphen is intended as a minus sign, so that this value is the maximum difference between breakover voltage in each direction.
- V_O Minimum output voltage.
- I_{TRM} Repetitive peak on-state current.
- I_{BO} Breakover current, usually the maximum required, and less than 20μA.

- I_R Maximum leakage current, usually less than 20μA.
- T_J Operating junction temperature, usually expressed as an acceptable range.

What Can Go Wrong

Like other semiconductors, a diac is heat sensitive. Usual precautions should be taken to allow sufficient ventilation and heat sinking, especially when components are moved from an open prototyping board to an enclosure in which crowding is likely.

Unexpected Triggering Caused by Heat

On a datasheet, a value for breakover current is valid only within a recommended temperature range. A buildup of heat can provoke unexpected triggering.

Low-Temperature Effects

A higher breakover voltage will be required by a diac operating at low temperatures, although the variation is unlikely to be greater than plus-or-minus 2% within a normal operating range. Temperature has a much more significant effect on a triac.

Manufacturing Tolerances

The breakover voltage for a diac is not adjustable, and may vary significantly between samples of the component that are supposed to be identical. The diac is not intended to be used as a precision component. In addition, while its breakover voltage should be the same in either direction, a difference of plus-or-minus 2% is possible (1% in some components).

triac



A **triac** is a gate-triggered type of *thyristor*. Its name was probably derived from the phrase “triode for AC,” and because it is not an acronym, it is not usually capitalized.

A thyristor is defined here as a semiconductor having four or more layers of p-type and n-type silicon. Because the thyristor predated integrated circuits, and in its basic form consists of a single multilayer semiconductor, it is categorized as a discrete component in this encyclopedia. When a thyristor is combined with other components in one package (as in a **solid-state relay**) it is considered to be an integrated circuit.

Other types of thyristor are the **SCR** (silicon-controlled rectifier) and the **diac**, each of which has its own entry in this encyclopedia.

Thyristor variants that are not so widely used, such as the *gate turn-off thyristor (GTO)* and *silicon-controlled switch (SCS)*, do not have entries here.

OTHER RELATED COMPONENTS

- **SCR** (see [Chapter 1](#))
- **diac** (see [Chapter 2](#))

What It Does

The **triac** is ubiquitous in AC dimmers for **incandescent lamps**. It is also used to control the speed of AC motors and the output of resistive heating elements. It is a type of *thyristor* which contains five segments of p-type and n-type silicon and has three leads, one of them attached to a gate that can switch a bidirectional flow of current between the other two. Its name was originally a trademark, generally thought to be derived from the phrase “triode for AC.” A triode was a common type of vacuum tube when thyristors were first introduced in the 1950s.

By comparison, a **diac** is a thyristor with only two leads, allowing current to flow in either direction when the component reaches a *breakover voltage*. Its name was probably derived from the

phrase “diode for AC.” It is often used in conjunction with a triac.

An **SCR** (silicon-controlled rectifier) is a thyristor that resembles a triac, as it has three leads, one of them a gate. However, it only allows current to flow in one direction.

Symbol Variants

The schematic symbol for a triac, shown in [Figure 3-1](#), resembles two diodes joined together, one of them inverted relative to the other. While a triac does not actually consist of two diodes, it is functionally similar, and can pass current in either direction.

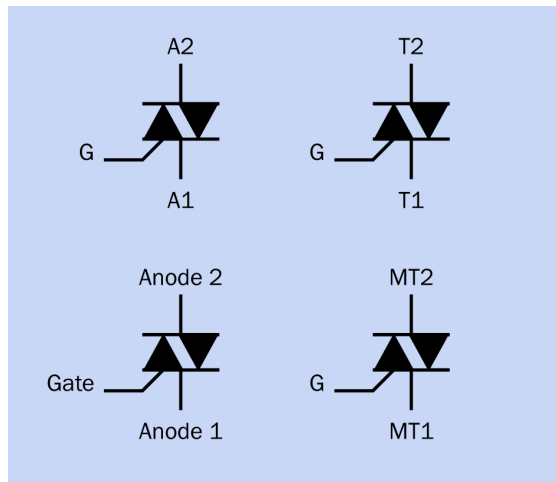


Figure 3-1. The schematic symbol for a triac, with four naming conventions that are used for its leads. The different conventions do not indicate any functional difference.

An appended bent line represents the gate. The labels for the other two leads are not standardized, and can be referred to as A1 and A2 (for Anode 1 and Anode 2), or T1 and T2 (for Terminal 1 and Terminal 2), or MT1 and MT2 (for Main Terminal 1 and Main Terminal 2). The choice of terms does not indicate any functional difference. In this encyclopedia entry, A1 and A2 are used.

The A1 terminal (or T1, or MT1) is always shown closer to the gate than A2 (or T2, or MT2). This distinction is important because although the triac can pass current in either direction, its behavior is somewhat asymmetrical.

- Voltages are expressed relative to terminal A1 (or T1, or MT1, if those terms are used).

The schematic symbol may be reflected or rotated, the black triangles may have open centers, and the placement of the bent line representing the gate may vary. However, terminal A1 is always nearer to the gate than terminal A2.

Figure 3-2 shows 12 of the 16 theoretical possibilities. All of these variants are functionally identical. Occasionally the symbol has a circle around it, but this style is now rare.

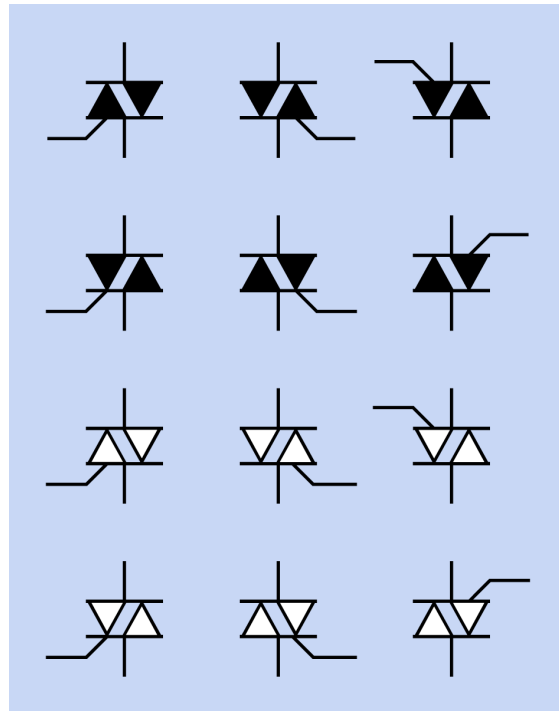


Figure 3-2. Interchangeable variants of the schematic symbol for a triac.

Triacs with various characteristics are shown in Figures 3-3, 3-4, and 3-5.

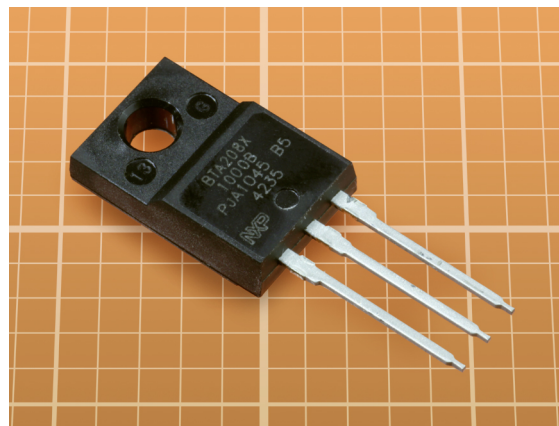


Figure 3-3. The BTA208X-1000B triac can conduct 8A continuous on-state current RMS, and withstands peak off-state voltage of up to 1,000V. This is a “snubberless” triac.

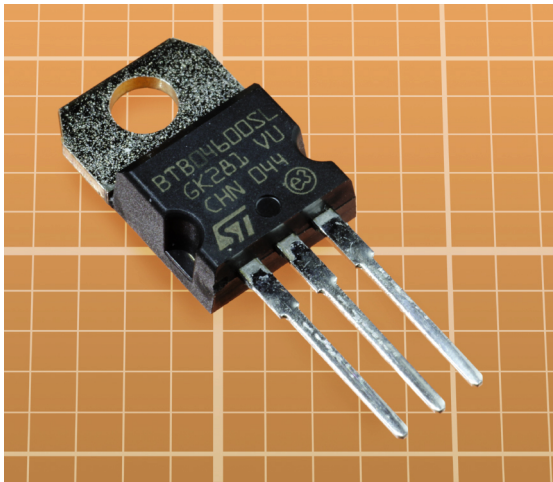


Figure 3-4. The BTB04-600SL triac can conduct 4A continuous on-state current RMS, and withstands peak off-state voltage of up to 600V.

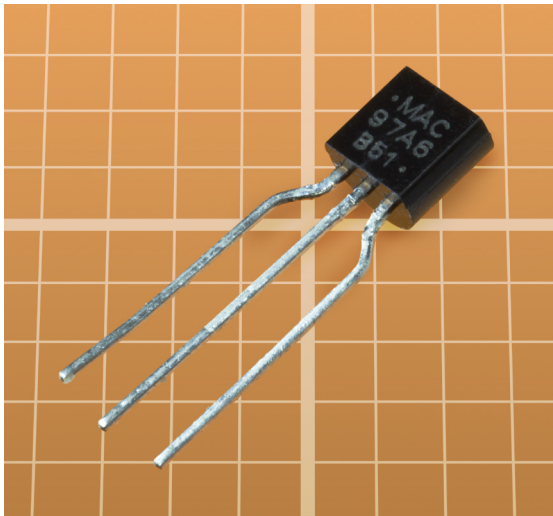


Figure 3-5. The MAC97A6 triac can conduct 0.8A continuous on-state current RMS, and withstands peak off-state voltage of up to 400V.

How It Works

When no gate voltage is applied, the triac remains in a passive state and will block current in either direction between A1 and A2, although a very small amount of *leakage* typically occurs. If the gate potential becomes sufficiently positive *or* negative relative to terminal A1, current can

begin to flow between A1 and A2 in *either* direction. This makes the triac ideal for controlling AC.

Quadrants

While a gate voltage is applied, four operating modes are possible. In each case, A1 is the reference (which can be thought of as being held at a neutral ground value). Because the triac is conducting AC, voltages above and below ground will occur. The four modes of operation are often referred to as four *quadrants*, and are typically arranged as shown in Figure 3-6.

In some reference sources (especially educational text books), current is shown with an arrow indicating a flow of electrons moving from negative to positive. Because the type of current flow is often undefined, diagrams should be interpreted carefully. In this encyclopedia, current is always shown flowing from a more-positive location to a more-negative location.

Quadrant 1 (upper right)

A2 is more positive than A1, and the gate is more positive than A1. Conventional current (positive to negative) will flow from A2 to A1. (This behavior is very similar to that of an **SCR**.)

Quadrant 2 (upper left)

A2 is more positive than A1, and the gate is more negative than A1. Once again, conventional current (positive to negative) will flow from A2 to A1.

Quadrant 3 (lower left)

A2 is more negative than A1, and the gate is more negative than A1. Conventional current is reversed from A1 to A2.

Quadrant 4 (lower right)

A2 is more negative than A1, but the gate is more positive than A1. Conventional current is reversed from A1 to A2.

- Note that two positive symbols or two negative symbols in Figure 3-6 do not mean that both locations are of equal voltage. They simply mean that these

locations are at potentials that are significantly different from A1.

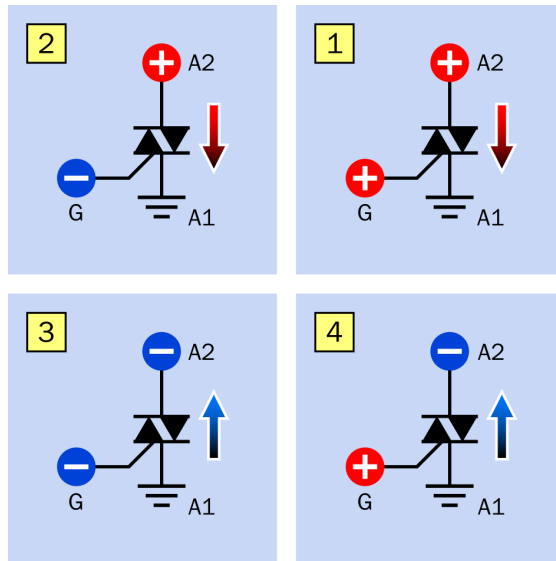


Figure 3-6. The “quadrants” of triac behavior. Positive and negative symbols indicate which terminal is “more positive” or “more negative” than A1. The ground symbol represents a potential midway between positive and negative. See text for details.

Suppose that gate current increases gradually. When it reaches the *gate threshold current* of the triac, the component starts conducting between A1 and A2. If the current between A1 and A2 rises above the value known as the *latching current*, it will continue to flow, even if gate current disappears completely.

If the self-sustaining current through the triac gradually diminishes, while there is no voltage applied to the gate, conduction between the main terminals will stop spontaneously when it falls below a level known as the *holding current*. This behavior is similar to that of an SCR. The triac now returns to its original state, blocking current until the gate triggers it again.

The triac is sufficiently sensitive to respond to rapid fluctuations, as in 50Hz or 60Hz AC.

Threshold, Latching, and Holding Current

Figure 3-7 shows the relationship between the gate threshold current, the latching current, and holding current. In the upper half of the figure, gate current is shown fluctuating until it crosses the threshold level. This establishes current flow between the main terminals, shown in the lower half of the figure. Prior to this moment, a very small amount of leakage current occurred (shown in the figure, but not to scale).

In this hypothetical scenario, the triac starts passing current between external components—and the current exceeds the latching level. Consequently, gate current can diminish to zero, and the triac remains conductive. However, when external factors cause the current between the main terminals to diminish below the holding level, the triac abruptly ceases to be conductive, and current falls back to the leakage level.

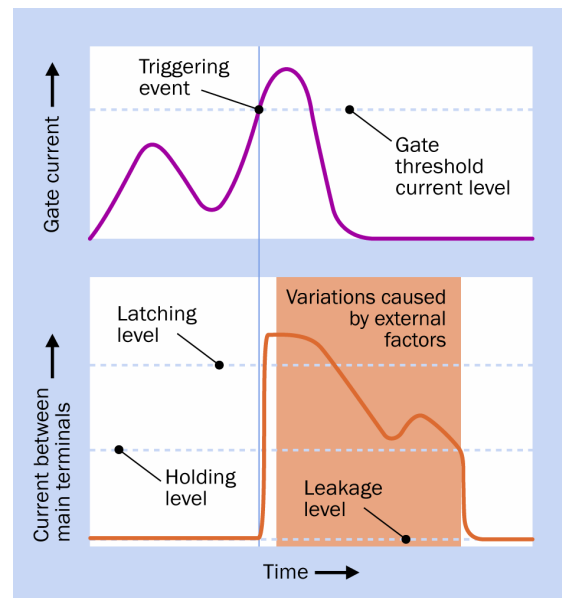


Figure 3-7. The relationship between gate current of a triac and the current between its main terminals. See text for details.

Unlike a **bipolar transistor**, a triac is either “on” or “off” and does not function as a *current ampli-*

fier. When it has been triggered, the impedance between A1 and A2 is low enough for heat dissipation to be manageable even at relatively high power levels.

Triac Testing

Figure 3-8 shows a circuit which can demonstrate the conductive behavior of a triac. For simplicity, this circuit is DC powered. In a real application, the triac is almost always used with AC.

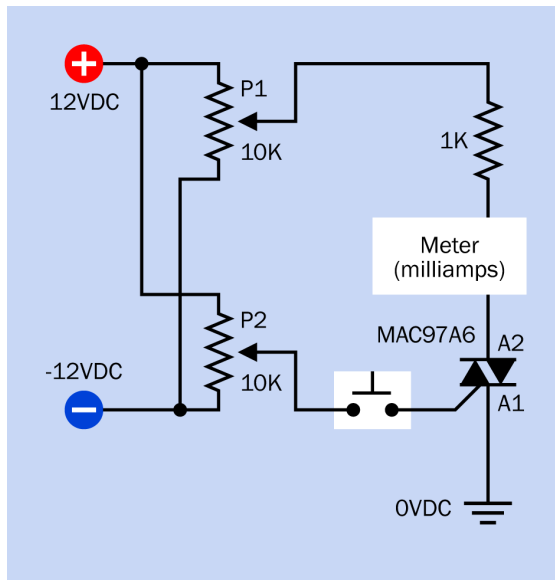


Figure 3-8. A test circuit to show the behavior of a triac when varying positive and negative potentials are applied to the gate and to the A2 terminal, relative to A1.

Note that this circuit requires at least a +12VDC and -12VDC power supply (higher values may also be used). The ground symbol represents a midpoint voltage of 0VDC, applied to terminal A1 of the triac, which is an MAC97A6 or similar. If a dual-voltage power supply is unavailable, the gate of the triac can be connected directly to +12VDC, omitting potentiometer P2; but in this case, only two operating modes of the triac can be demonstrated by turning potentiometer P1.

Each potentiometer functions as a voltage divider between the positive and negative sides of the power supply. P1 applies a positive or negative

voltage to A2, relative to A1. P2 applies a positive or negative voltage to the gate, relative to A1.

If the test begins with both potentiometers at the top ends of their range, A1 and G both have a positive potential relative to A1, so that the triac is now in quadrant 1 of its operating modes. Pressing the pushbutton should cause it to start conducting current limited by the 1K resistor, and the meter should change from measuring 0mA to around 12mA. If the pushbutton is released, the triac should continue to conduct current, because 12mA is above this triac's latching level. If P1 is slowly moved toward the center of its range, the current diminishes, ceasing when it falls below the holding level. If P1 is now moved back to the top of its range, the current will not resume until the triac is retriggered with the pushbutton.

The test can be repeated with P1 at the top of its range and P2 at the bottom of its range, to operate the triac in quadrant 2; P1 at the bottom of its range and P2 at the bottom of its range, to operate the triac in quadrant 3; and P1 at the bottom of its range and P2 at the top of its range, to operate the triac in quadrant 4. The functionality should be the same in each case. The pushbutton will initiate a flow of current, which will diminish when P1 is turned toward the center of its range.

In any of these quadrants, P2 can be turned slowly toward the center of its range while the pushbutton is pressed repeatedly. This will allow empirical determination of the gate threshold current for this triac. The meter, measuring milliamps, will measure the current if it is inserted between the wiper of the potentiometer and the gate of the triac.

The test circuit is shown installed on a breadboard in Figure 3-9. In this photograph, the red and blue wires at left supply +12VDC and -12VDC relative to the black ground wire at top right. The yellow and green wires connect with a meter set to measure milliamps. The red button is a tactile switch, while the MAC97A6 triac is just above it

and to the left. The square blue 10K trimmers are set to opposite ends of their scales, so that the meter will show current flowing when the tactile switch is pressed.

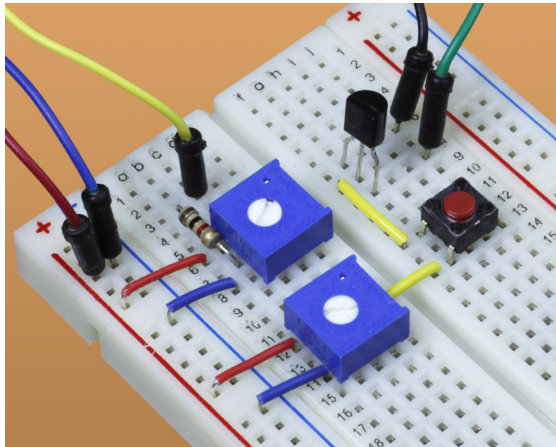


Figure 3-9. A breadboarded triac test circuit.

Breakover Voltage

If a much higher voltage is applied to A2, the triac can be forced to conduct current without any triggering voltage being applied to the gate. This occurs when the potential between A1 and A2 reaches the triac's *breakover voltage*, although the component is not designed to be used this way. The concept is illustrated in [Figure 3-10](#), which can be compared with the behavior of an SCR illustrated in [Figure 1-8](#) and the behavior of a diac shown in [Figure 2-5](#). While the term *breakdown voltage* defines the minimum reverse voltage required to force a diode to conduct, *breakover voltage* refers to the minimum forward voltage that has this effect. Because a triac is designed to conduct in both directions, it can be thought of as having a breakover voltage in both directions.

In [Figure 3-10](#), the numbers in yellow squares are the quadrants referred to in [Figure 3-6](#). The solid curve represents current flow if a triggering voltage is applied to the gate while a positive or negative potential is applied to A2, relative to A1. If the gate is not triggered while the voltage be-

tween A1 and A2 gradually increases, the dashed section of the curve illustrates the outcome when the component reaches breakover voltage. Although this may not damage the triac, the component becomes uncontrollable.

- In normal usage, the voltage between A1 and A2 should not be allowed to reach breakover level.

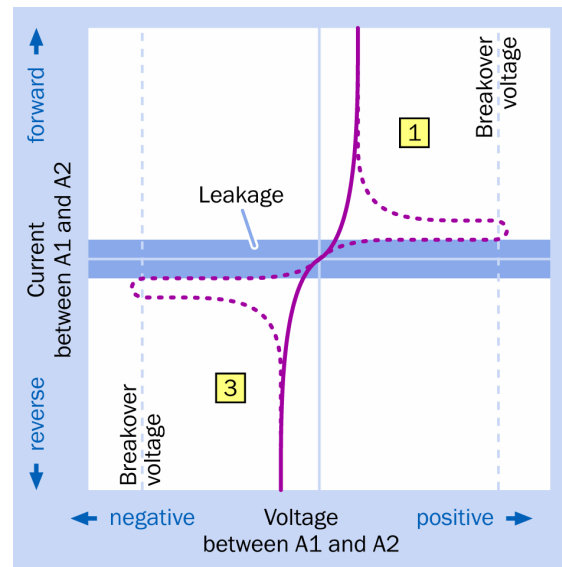


Figure 3-10. The solid curve shows current passing between A1 and A2 in a hypothetical triac, for varying voltages, while triggering voltage is applied to the gate. The dashed curve assumes that no triggering voltage is applied to the gate. The numbers in yellow squares are the quadrants of triac operation.

Switching AC

"Switching" AC with a triac means interrupting each pulse of current so that only a portion of it is conducted through to the load. Usually this is done with the triac functioning in quadrants 1 and 3. In quadrant 3, the polarity of the flow between A1 and A2 is opposite to that in quadrant 1, and the gate voltage is also reversed. This enables a relatively simple circuit to control the duration of each half-cycle passing through the triac. The theory of this circuit is shown in [Figure 3-11](#).

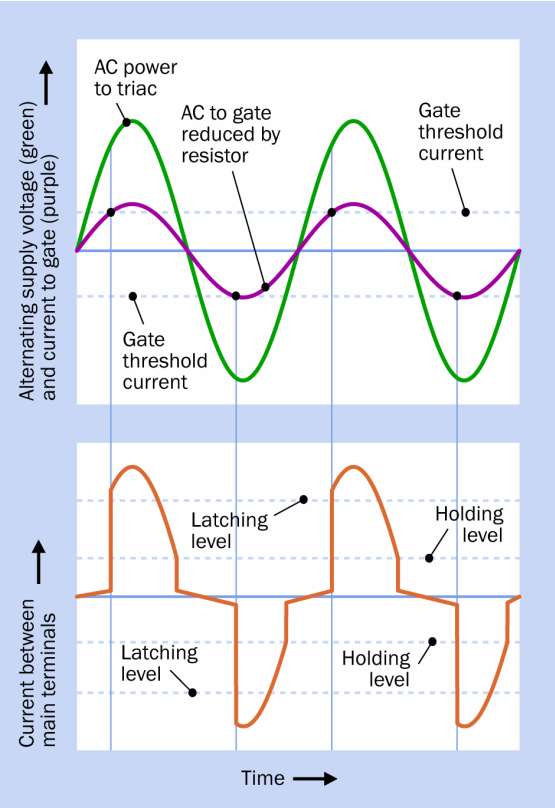


Figure 3-11. To moderate the power of AC current, a triac can block a section of each AC pulse.

The upper section of [Figure 3-11](#) shows alternating voltage to the triac in green. The purple curve represents the gate current of the triac, reduced by a variable resistor. (The figure is for conceptual purposes only; the alternating power supply voltage and the fluctuating gate current cannot actually share the same vertical scale of a graph.)

[Figure 3-11](#) can be compared with [Figure 3-7](#), except that the negative threshold level for the gate is now shown as well as the positive threshold level. Remember, either a positive or negative voltage can activate the gate.

In [Figure 3-11](#), initially the triac is nonconductive. As time passes, the gate current reaches the threshold level, and this triggering event enables current to flow between the main terminals of the triac, as shown in the lower part of the figure.

This current exceeds the latching level, so it continues to flow, even though the gate current diminishes below its threshold level. Finally the current between the main terminals falls below the holding level, at which point the triac stops conducting. It waits for the next triggering event, which occurs as the power supply swings to negative.

This simple system blocks a section of each AC pulse, which will vary in length depending how much current is allowed to flow through the gate. Because the blocking process occurs rapidly, we notice only the reduced overall power passing through the triac (in terms of the brightness of a light, the heat emitted by a resistive element, or the speed of a motor).

Unfortunately, there is a problem in this scenario: the triac does not quite behave symmetrically. Its gate threshold level for positive current is not exactly equal and opposite to its gate threshold level for negative current. The upper part of [Figure 3-11](#) shows this flaw in the differing vertical offsets of the positive and negative thresholds from the central zero line.

The result is that negative AC pulses through the triac are shorter than positive pulses. This asymmetry produces harmonics and noise that can feed back into power supply wiring, interfering with other electronic equipment. The actual disparities in gate response, in each quadrant of operation for two triacs, are shown in [Figure 3-12](#).

Current capacity of triac	Ratio of gate current required to conduct, relative to quadrant 1		
	Quadrant 2	Quadrant 3	Quadrant 4
4 amp	1.6	2.5	2.7
10 amp	1.5	1.4	3.1

Figure 3-12. Because the internal structure of a triac is asymmetrical, it requires a different trigger current in each of its operating quadrants. This table, derived from a Littelfuse technical briefing document, shows the ratio of the minimum trigger current in quadrants 2, 3, and 4 relative to quadrant 1.

See [Figure 1-14](#) for a graph illustrating phase control in the SCR. See [“Phase Control”](#) for a discussion of phase in AC waveforms generally.

Triac Triggered by a Diac

The problem of asymmetrical triggering can be overcome if the triac is triggered with a voltage pulse generated by another component that does behave symmetrically. The other component is almost always a diac, which is another type of thyristor. Unlike an SCR or a triac, it has no gate. It is designed to be pushed beyond its breakover voltage, at which point it latches and will continue to conduct until current flowing through it diminishes below its holding level. See [Chapter 2](#) for more information about the diac.

In [Figure 3-13](#), the diac is shown to the right of the triac, and is driven by a simple RC network consisting of a fixed resistor, a potentiometer, and a capacitor. (In an actual application, the RC network may be slightly more complex.) The capacitor takes a small amount of time to charge during each half-cycle of AC. The length of this delay is adjusted by the potentiometer, and determines the point in each AC half-cycle when the voltage to the diac reaches breakover level. Because the delay affects the phase of the AC, this adjustment is known as [phase control](#).

As the voltage exceeds breakover level, the diac starts to pass current through to the gate of the triac, and triggers it. The holding level of the diac is lower than its latching level, so it continues to pass current while the capacitor discharges and the voltage diminishes. When the current falls below the holding level, the diac stops conducting, ready for the next cycle. Meanwhile, the triac continues to pass current until the AC voltage dips below its holding level. At this point, the triac becomes nonconductive until it is triggered again.

This chopped waveform will still create some harmonics, which are suppressed by the coil and capacitor at the left side of the circuit in [Figure 3-13](#).

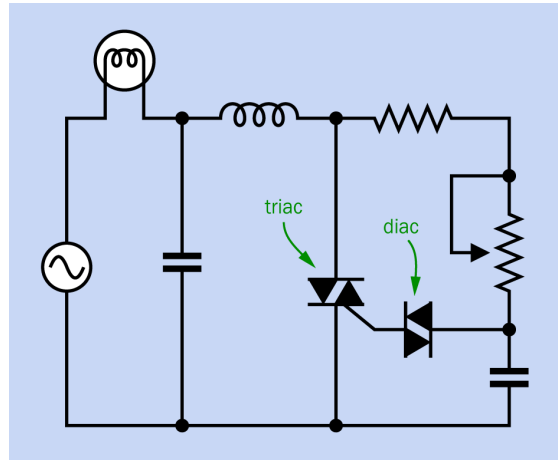


Figure 3-13. A minimal schematic showing typical operation of a triac, with a diac supplying pulses to the triac gate. The potentiometer adjusts the delay created by the capacitor.

Other Triac Drivers

It is possible, although unusual, to drive a triac from a source other than a diac.

Simple on-off control can be achieved by using a special optocoupler such as the MOC3162 by Fairchild Semiconductor. This emits a switching signal to a triac only when the AC voltage passes through zero. A [zero cross](#) circuit is desirable because it creates much less interference. The use of an optocoupler helps to isolate the triac from other components.(((“zero cross circuit”)))

Phase control can be achieved using an optocoupler such as the H11L1, which can be driven by rectified but unsmoothed AC after it passes through a Zener diode to limit the voltage. The output from the optocoupler is logic-compatible and can be connected with the input to a timer such as the 555, set to one-shot mode. Each pulse from the timer passes through another optocoupler such as the MOC3023, which uses an internal LED to trigger the gate of a triac.

Yet another possibility is to use the programmed output from a microcontroller, through an optocoupler, to control the gate of a triac. An online

search for the terms “microcontroller” and “triac” will provide some additional suggestions.

Charge Storage

While switching AC, the internal charge between A1 and A2 inside the triac requires time to dissipate before the reverse voltage is applied; otherwise, *charge storage* occurs, and the component may start to conduct continuously. For this reason, the triac is normally restricted to relatively low frequencies such as domestic 60Hz AC power.

When a triac controls a motor, the phase lag between voltage and current associated with an inductive load can interfere with the triac’s need for a transitional moment between a positive and negative voltage cycle. In a datasheet, the term *commutating dv/dt* defines the rate of rise of opposite polarity voltage that the triac can withstand without locking into a continuous-on state.

An RC *snubber network* is often wired in parallel with A1 and A2 to control the rise time of voltage to the triac, as shown within the darker blue rectangle in Figure 3-14, where a resistor and capacitor have been added just to the left of the triac. The highest resistance and lowest capacitance, consistent with trouble-free operation, should be chosen. Typical values are 47Ω to 100Ω for the resistor, and 0.01μF to 0.1μF for the capacitor.

Variants

Triacs are available in through-hole and surface-mount packages.

Some components that are referred to as triacs actually contain two SCR components of opposite polarity. The “alternistor” range from Littelfuse is an example. The SCR will tolerate faster voltage rise times than a conventional triac, and is more suitable for driving inductive loads such as large motors.

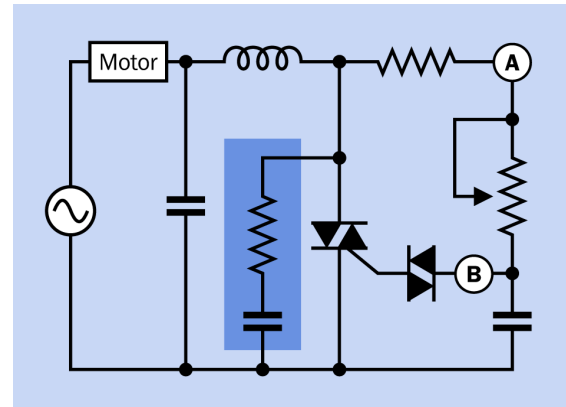


Figure 3-14. To prevent a triac from locking itself into a continuous-on state while driving an inductive load such as a motor, a snubber circuit can be added (shown here as a resistor and capacitor in the darker blue rectangle to the left of the triac).

A *snubberless triac*, as its name implies, is designed to drive an inductive load without need for a snubber circuit. An example is the STMicroelectronics BTA24. Datasheets for this type of component impose some limits that may be stricter than for a generic triac.

Values

Surface-mount triacs are typically rated between 2A to 25A of switched AC current (RMS), the higher-current versions being as large as 10mm square. The necessary gate trigger voltage may range from 0.7V to 1.5V. Through-hole packages may be capable of slightly higher currents (up to 40A), with gate trigger voltages of 1V to 2.5V being common.

As noted previously, the majority of triacs are restricted to relatively low frequency switching, 60Hz being very common.

Abbreviations in datasheets are likely to include:

- V_{DRM} or V_{RRM} Peak repetitive reverse off-state voltage. The maximum reverse voltage that the component will withstand in its “off” state without experiencing damage or allowing current to pass.

- V_{TM} The maximum voltage difference between A1 and A2, measured with a short pulse width and low duty cycle.
- V_{GT} Gate trigger voltage necessary to produce the gate trigger current.
- I_{DRM} Peak repetitive blocking current (i.e., maximum leakage).
- I_{GM} Maximum gate current.
- I_{GT} Minimum gate trigger current.
- I_H Holding current.
- I_L Latching current.
- $I_{T(RMS)}$ On-state RMS current. The maximum value passing through the component on a continuous basis.
- I_{TSM} Maximum non-repetitive surge current. Specified at a stated pulse width, usually 60 Hz.
- T_C Case temperature, usually expressed as an acceptable range.
- T_J Operating junction temperature, usually expressed as an acceptable range.

What Can Go Wrong

Like other semiconductors, a triac is heat sensitive. Usual precautions should be taken to allow sufficient ventilation and heat sinking, especially when components are moved from an open prototyping board to an enclosure in which crowding is likely.

Unexpected Triggering Caused by Heat

On a datasheet, a value for triggering current is valid only within a recommended temperature range. A buildup of heat can provoke unexpected triggering.

Low-Temperature Effects

Significantly higher gate current will be required by a triac operating at low temperatures. It is quite possible that the component will need twice as much current at 25° C compared with 100° C, junction temperature. If the triac receives insufficient current, it will not turn on.

Wrong Type of Load

If an **incandescent lamp** is replaced with a **fluorescent light** or **LED area lighting**, a pre-existing triac may no longer work as a dimmer. Fluorescent lamps will have some inductance, and may also provide a capacitive load, either of which will interfere with the normal behavior of a triac.

The light output of an LED varies very differently compared with the light output of an incandescent bulb, in response to reduction in power. Therefore an LED should be dimmed using pulse-width modulation that is appropriate for its output characteristics. A triac is generally not suitable.

Wrongly Identified Terminals

A triac is often thought of as a symmetrical device, because it is designed to switch AC current using either positive or negative voltage at the gate. In reality, its behavior is asymmetrical, and if it is installed “the wrong way around” it may function erratically or not at all.

Failure to Switch Off

As already noted (see [“Charge Storage” on page 23](#)), a triac will tend to suffer from *charge storage* if there is insufficient time between the end of one half-cycle and the beginning of the next. A component that works with a resistive load may cease to function if it is used, instead, to power an inductive load.

solid-state relay

A **solid-state relay** is less-commonly referred to by its acronym, *SSR*. It is sometimes regarded as an **optocoupler**, but in this encyclopedia the two components have separate entries. An optocoupler is a relatively simple device consisting of a light source (usually an **LED**) and a light sensor, in one package. It is used primarily for isolation rather than to switch a high current. A solid-state relay can be thought of as a substitute for an electromagnetic **relay**, usually has additional components in its package, and is intended to switch currents of at least 1A.

A component that works like a solid-state relay but only switches a 5V (or lower) logic signal may be referred to as a *switch*, even though it is entirely solid-state. This type of component is included in this entry because it functions so similarly to a solid-state relay.

OTHER RELATED COMPONENTS

- **electromagnetic relay** (see Volume 1)
- **optocoupler** (see [Chapter 5](#))

What It Does

A **solid-state relay** (SSR) is a semiconductor package that emulates an electromagnetic **re-lay** (see Volume 1). It switches power on or off between its output terminals in response to a smaller current and voltage between its input terminals. Variants can switch AC or DC and may be controlled by AC or DC. An SSR functions as a *SPST switch*, and is available in normally open or normally closed versions. SSRs that function as an SPDT switch are relatively unusual and actually contain more than one SSR.

No single schematic symbol has been adopted to represent a solid-state relay, but some alternatives are shown in [Figure 4-1](#):

Top

An unusually detailed depiction of an SSR that switches DC current using MOSFETs. Symbols for this device often omit the diodes

on the output side and may simplify the MOSFET symbols.

Bottom left

An SSR that uses an internal **triac** to switch AC. The box labeled 0x indicates that this is a *zero-crossing* relay, meaning that it switches when alternating voltage crosses the 0V level from positive to negative or negative to positive.

Bottom right

A generic SSR, showing a symbol for a normally open relay, although whether it is designed for AC or DC is unclear.

Advantages

- Great reliability and long life.
- No physical contacts that are vulnerable to arcing and erosion or (under extreme conditions) that could weld themselves together.

- Very fast response, typically $1\mu\text{s}$ on and $0.5\mu\text{s}$ off.
- Very low power consumption on the input side, as low as 5mA at 5VDC . Many solid-state relays can be driven directly from logic chips.
- Lack of mechanical noise.
- No contact bounce; a clean output signal.
- No coil that would introduce back EMF into the circuit.
- Safe with flammable vapors, as there is no sparking of contacts.
- Often smaller than a comparable electromagnetic relay.
- Insensitive to vibration.
- Safer for switching high voltages, as there is complete internal separation between input and output.
- Some variants work with input control voltages as low as 1.5VDC . Electromagnetic relays typically require at least 3VDC (or more, where larger relays are required to switch higher currents).

- More vulnerable than an electromagnetic relay to surges and spikes in the current that is switched on the output side.

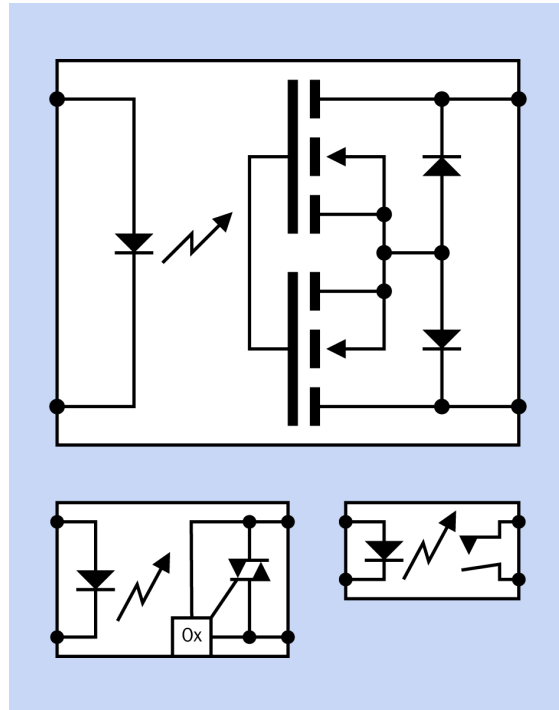


Figure 4-1. Schematic symbols for solid-state relays have not been standardized. See text for details.

Disadvantages

- Less efficient; its internal impedance introduces a fixed-value voltage drop on the output side (although this may be negligible when switching higher voltages).
- Generates waste heat in its “on” mode, in accordance with the voltage drop.
- Passes some *leakage* current (usually measured in microamps) on the output side when the relay is supposed to be “off.”
- A DC solid-state relay usually requires observation of polarity on the output side. An electromagnetic relay does not.
- Brief voltage spikes on the input side, which would be ignored by a slower electromagnetic relay, may trigger a solid-state relay.

How It Works

Almost all modern SSRs contain an internal **LED** (light-emitting diode, see [Chapter 22](#)) which is switched on by the control input. Infrared light from the LED is detected by a sensor consisting of one or more **phototransistors** or **photodiodes**. In a relay that controls DC current, the sensor usually switches a **MOSFET** (see Volume 1) or an **SCR** (silicon-controlled rectifier—see [Chapter 1](#)). In relays that control AC current, a **triac** (see [Chapter 3](#)) controls the output. Because the input side and the output side of the SSR are linked only by a light signal, they are electrically isolated from each other.

The MOSFETs require so little power, it can be provided entirely by light falling on an array of 20 or more photodiodes inside the SSR package.

Typical solid-state relays are shown in Figures 4-2 and 4-3.



Figure 4-2. A solid-state relay capable of switching up to 7A DC. See text for a detailed description.

The Crydom DC60S7 accepts a control voltage ranging from 3.5VDC to 32VDC, with a typical input current of less than 3mA. Maximum turn-on time is 0.1ms and maximum turn-off time is 0.3ms. This relay can switch up to 7A and tolerates a surge of up to twice that current. It imposes a voltage drop of as much as 1.7VDC, which can become a drawback when switching voltages that are significantly lower than its maximum 60VDC. The electronics are sealed in thermally conductive epoxy, mounted on a metal plate approximately 1/8" thick which can be screwed down onto an additional heat sink.

The Crydom CMX60D10 tolerates a more limited range of control voltages (3VDC to 10VDC) and requires a higher input current of 15mA at 5VDC. However, its very low maximum on-state resistance of 0.018Ω imposes a much smaller voltage drop of less than 0.2 volts when passing 10A. This results in less waste heat and enables a *single-inline package* (SIP) without a heat sink. The

CMX60D10 weighs 0.4 ounces, as opposed to the 3 ounces of the DC60S7. Relays from other manufacturers use similar packaging and have similar specifications.



Figure 4-3. A solid-state relay capable of switching up to 10A. Its lower internal resistance results in less waste heat and enables a smaller package. See text for a detailed description.

Variants

Many solid-state relays have protective components built into the package, such as a varistor on the output side to absorb transients. Check datasheets carefully to determine how much protection from external components may be necessary when switching an inductive load.

Instantaneous versus Zero Crossing

A *zero crossing* SSR is one that (a) switches AC current and (b) will not switch “on” until the instant when the AC voltage crosses through 0V. The advantages of this type are that it does not have to be built to switch such a high current, and creates minimal voltage spike when the switching occurs.

All SSRs that are designed to switch AC will wait for the next voltage zero crossing before switching to their “off” state.

NC and NO Modes

Solid-state relays are SPST devices, but different models may have a normally closed or normally open output. If you require double-throw operation, two relays can be combined, one normally closed, the other normally open. See [Figure 4-4](#). A few manufacturers combine a normally closed relay and a normally open relay in one package, to emulate a SPDT relay.

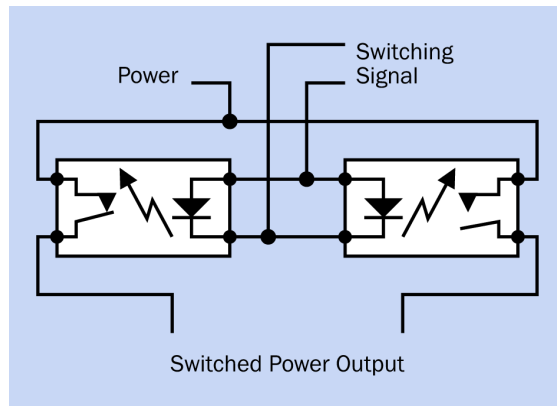


Figure 4-4. A normally closed solid-state relay can be paired with a normally open solid-state relay to emulate a SPDT switch. This combination is available in a single package from some manufacturers.

Packaging

High-current solid-state relays are often packaged with screw terminals and a metal base that is appropriate for mating with a [heat sink](#). Some are sold with heat sinks integrated. Spade terminals and crimp terminals may be optional. The Crydom DC60S7 shown in [Figure 4-2](#) is an example. This type of package may be referred to as [industrial mount](#).

Lower-current solid-state relays (5A or less), and those with a very low output resistance, may be packaged with single-inline pins for through-hole mounting in circuit boards.

Solid-State Analog Switch

DIP packaging may be used for solid-state relays that are designed for compatibility with the low voltages and currents of logic chips. This type of

component may be referred to simply as a [switch](#). The 74HC4316 is an example, pictured in [Figure 4-5](#).

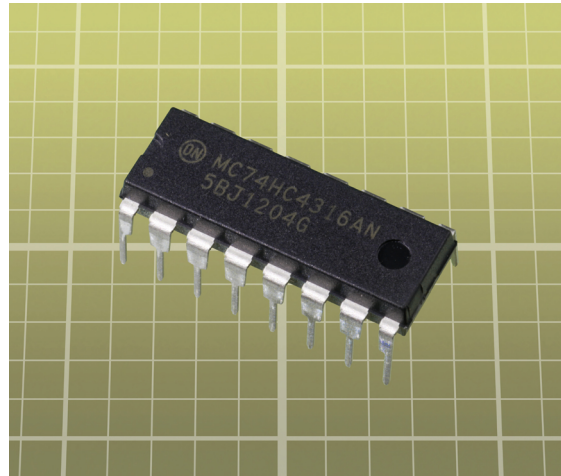


Figure 4-5. This DIP package contains four “switches” that function as solid-state relays but are restricted to low voltages and currents, compatible with logic chips. See text for details.

Typically the control voltage and the switched voltage are limited between +7V and –7V, with a maximum output current of 25mA. Each internal switch has its own Control pin, while an additional Enable pin forces all switches into an “off” state if its logic state is high. The simplified functionality of this component is illustrated in [Figure 4-6](#), without showing internal optical isolation.

The “on” resistance of each internal pathway will be approximately 200Ω when the component is powered with +5VDC on the positive side and 0VDC on the negative side. This resistance drops to 100Ω if the negative power supply is –5VDC.

If all of the outputs from the chip are shorted together, it functions as a **multiplexer** (see [Chapter 16](#)). In fact, this type of switch component is often listed in catalogs as a multiplexer, even though it has other applications.

Because the component tolerates equal and opposite input voltages, it is capable of switching AC.

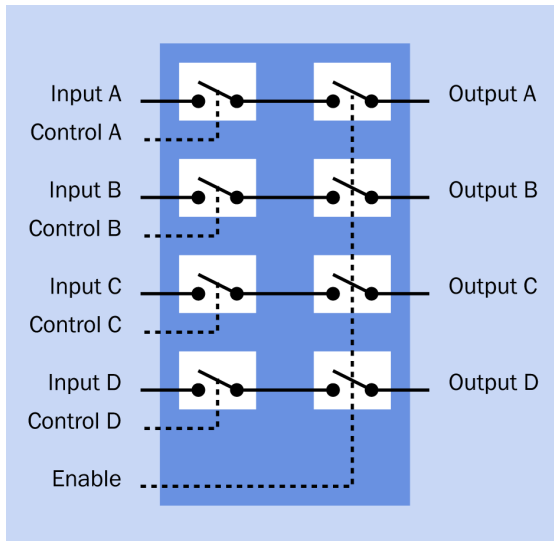


Figure 4-6. The functionality of a chip containing four solid-state analog switches. A high state on a Control pin closes its associated switch. The Enable pin must be held low for normal operation; a high Enable state forces all the switches into the “off” position. If the outputs are tied together, this component can function as a multiplexer.

Values

Industrial-mount solid-state relays typically can switch currents ranging from 5A to 500A, with 50A being very common. The higher-current relays mostly require DC control voltage; 4V to 32V are typical, although some versions can go much higher. They contain an SCR or triac to switch AC.

Smaller solid-state relays in SIP, DIP, or surface-mount packages often use MOSFETs on the output side, and are often capable of switching up to 2A or 3A. Some can switch either AC or DC, depending on the way the output is wired. The LED on the input side may require as little as 3mA to 5mA for triggering.

How to Use It

Solid-state relays find their primary uses in telecommunications equipment, industrial control systems and signalling, and security systems.

The component is very simple externally. Power on the input side can come from any source capable of delivering the voltage and current specified by the manufacturer, and any device that doesn’t exceed maximum current rating can be connected to the output side, so long as provision is made for suppressing back-EMF from an inductive load, as shown in [Figure 4-7](#). Often a solid-state relay can be substituted directly for an electromagnetic relay, without modifying the circuit.

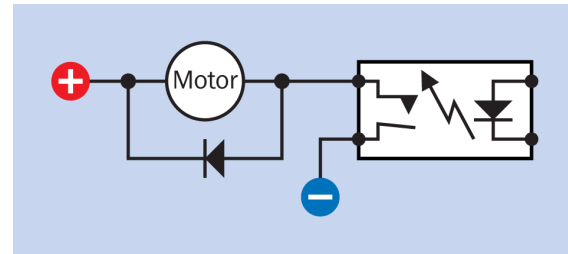


Figure 4-7. Use of a diode around an inductive load, to protect a solid-state relay from back-EMF.

Solid-state relays are heat sensitive, and their rating for switching current will diminish as their temperature increases. Manufacturer datasheets will provide specific guidance. Using a heat sink will greatly improve the performance. Bear in mind that the relay generates heat continuously while it is in its “on” mode—about 1 watt per ampere.

Because it requires so little current on the input side (typically no more than 15mA), a solid-state relay can usually be driven directly by chips such as [microcontrollers](#) that would not be able to activate an equivalent electromagnetic relay.

Applications may take advantage of the solid-state relay’s reliability, immunity to vibration, lack of contact sparking, freedom from coil-induced surges on the input side, and lack of contact bounce on the output side. A solid-state relay is ideal within digital equipment that is sensitive to power spikes. It may switch a fuel pump that handles volatile, flammable liquids, or a wastewater pump in a basement subject to

flooding (where long-term zero-maintenance reliability is necessary, and contact corrosion could be a risk in electromagnetic relays). Small solid-state relays can switch motors in robots or appliances where vibration is common, and are often used in arcade games.

What Can Go Wrong

Overheating Caused by Overloading

Relays must be *derated* when used at operating temperatures above the typical 20 or 25° C for which their specification applies. In other words, the sustained operating current must be reduced, usually by an amount such as 20% to 30% for each 10-degree increase in ambient temperature. Failure to observe this rule may result in failure of the component. Burnout may also occur if a high-current solid-state relay is used without a heat sink, or the heat sink isn't big enough, or *thermal compound* is not applied between the solid-state relay and the heat sink.

Overheating Caused by Bad Terminal Contact

If the screw terminals on the output side of a high-current solid-state relay are not tightened sufficiently, or if there is a loose spade terminal, or if a crimped connection isn't crimped tightly enough, the poor contact will create electrical resistance, and at high currents, the resistance will create heat, which can cause the solid-state relay to overheat and burn out.

Overheating Caused by Changing Duty Cycle

If a high-current solid-state relay is chosen for an application where it is in its "on" state only half the time, but the application changes during product development so that the solid-state relay is in its "on" state almost all the time, it will have to dissipate almost twice as much heat. Any time the duty cycle is changed, heat should be considered. The possibility of the relay being

used in an unconventional or unexpected manner should also be considered.

Overheating Caused by Component Crowding

Overheating increases dramatically when components are tightly crowded. At least 2cm (3/4") should be allowed between components.

Overheating in Dual Packaging

When a package contains two solid-state relays, the additive effects of the heat created by each of them must be considered.

Reverse-Voltage Burnout

Because a solid-state relay is more sensitive to *back-EMF* than an electromagnetic relay, greater care should be used to protect it from reverse voltage when switching inductive loads. A *protection diode* should be used, and a *snubber* can be added between its output terminals, if it is not included inside the relay package.

Low Voltage Output Current May Not Work

Unlike electromagnetic relays, solid-state relays require some voltage on the output side to enable their internal operation. If there is no voltage, or only a very low voltage, the SSR may not respond to an input. The minimum voltage required on the output side is usually specified in a datasheet.

To test a solid-state relay, apply actual voltages on input and output sides and use a load such as an incandescent light bulb. Merely applying a meter on the output side, set to measure continuity, may not provide sufficient voltage to enable the relay to function, creating the erroneous impression that it has failed.

Inability to Measure AC Output

When a multimeter is used to test continuity across the output of an AC-switching solid-state relay of zero-crossing specification, the meter will generate enough voltage to prevent the

solid-state relay from finding zero voltage across its output terminals, and consequently the solid-state relay won't switch its output.

Relay Turns On but Won't Turn Off

When a solid-state relay controls a relatively high-impedance load such as a small **solenoid** (see Volume 1) or a **neon bulb** (see [Chapter 19](#)), the relay may switch the device on but will seem unable to switch it off. This is because the leakage current of the solid-state relay, in its "off" state, may be just enough to maintain the load in its "on" state.

If an SSR containing a triac is used erroneously to switch DC, it will not be able to switch off the current.

Relays in Parallel Won't Work

Two solid-state relays usually cannot be used in parallel to switch twice as much current. Because of small manufacturing variances, one relay will switch on a moment before the other. When the first relay is on, it will divert the load current away from the second relay. The second relay needs a small amount of current on its output side, to function. Without any current, it will not switch

on. This means the first relay will pass the total current without any help from the second relay, and will probably burn out, while the second relay does nothing.

Output Device Doesn't Run at Full Power

A solid-state relay imposes a voltage reduction on its output side. This will be a fixed amount, not a percentage. When switching 110V, this difference may be negligible; when switching 12V, it may deliver only 10.5V, which represents enough of a drop to cause a motor or a pump to run noticeably more slowly. The internal switching device inside the relay (MOSFET, triac, SSR, or bipolar transistor) will largely determine the voltage drop. Check the manufacturer's datasheet before using the relay.

Solid-State Relays and Safety Disconnects

Because a solid-state relay always allows some leakage in its "off" state, it can still deliver a shock when used to switch high voltages. For this reason, it may not be suitable in a safety disconnect.

optocoupler

Sometimes known as an *optoelectronic coupler*, *opto-isolator*, *photocoupler*, or *optical isolator*.

A **solid-state relay** is sometimes referred to as an **optocoupler**, but in this encyclopedia it has a separate entry. An optocoupler is a relatively simple device consisting of a light source (usually an LED) and a light sensor, both embedded in one package. It is used primarily for isolation rather than to switch a high current. A solid-state relay can be thought of as a substitute for an electromagnetic **relay**, usually has additional components in its package, and is intended to switch currents of at least 1 A.

OTHER RELATED COMPONENTS

- **electromagnetic relay** (see Volume 1)
- **solid-state relay** (see [Chapter 4](#))

What It Does

An optocoupler allows one section of a circuit to be electrically isolated from another. It protects sensitive components, such as logic chips or a microcontroller, from voltage spikes or incompatible voltages in other sections of a circuit. Optocouplers are also used in medical devices where a patient has to be protected from any risk of electric shock, and are used in devices which conform with the MIDI standard for digital control of music components.

In [Figure 5-1](#), three possible applications for an optocoupler are suggested:

Top

The output from a logic chip passes through an optocoupler to an inductive load such as a relay coil, which may create voltage spikes that would be hazardous to the chip.

Center

The noisy signal from an electromagnetic switch passes through an optocoupler to the input of a logic chip.

Bottom

The low-voltage output from a sensing device on a human patient passes through an optocoupler to some medical equipment, such as an EEG machine, where higher voltages are used.

Internally, an optocoupler works on the same principle as a **solid-state relay**. An LED is embedded on the input side, shining light through an interior channel or transparent window to a sensing component that is embedded on the output side. Because the only internal connection is a light beam, the input and output of the optocoupler are isolated from each other.

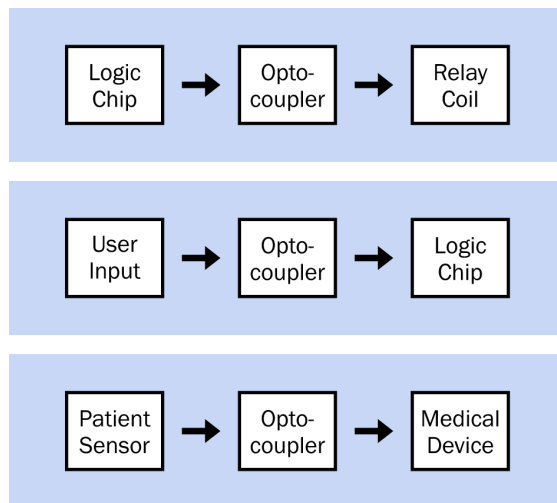


Figure 5-1. Possible applications for a photocoupler. See text for details.

Isolation transformers were traditionally used for this purpose prior to the 1970s, when optocouplers became competitive. In addition to being smaller and cheaper, an optocoupler can also pass slow-changing signals or on-off DC states which a transformer would ignore.

More recently, inductive and capacitive coupling components have become available in surface-mount packages that are competitive with optocouplers for high-speed data transfer. They also claim to be more durable. Because of the gradual reduction in output from an LED, the performance of an optocoupler degrades over time, and is typically rated for up to 10 years.

How It Works

The LED in an optocoupler almost always emits light in the near-infrared part of the spectrum, and is matched to the sensitivity of a **photo-transistor**, or a **photodiode**, or (less often) a **photoresistor** that provides the output. Photosensitive **triacs** and **SCRs** are also sometimes used.

The most common type of optocoupler uses a bipolar phototransistor with an open-collector

output. Schematic symbols for this type are shown in [Figure 5-2](#):

Top left

The most common generic form.

Top right

Two diodes on the input side allow the use of alternating current.

Center left

An additional terminal allows addition of bias to the photosensitive base of the output transistor, to reduce its sensitivity.

Center right

An Enable signal can be used as the input to the NAND, suppressing or enabling the output.

Bottom left

A *photodarlington* allows higher emitter current.

Bottom right

Relatively uncommon, and is also used for a **solid-state relay**.

In each symbol, the diode is an LED, and the zig-zag arrow indicates light that is emitted from it. A pair of straight arrows, or wavy arrows, may alternatively be used.

An optocoupler in through-hole DIP format is shown in [Figure 5-3](#).

An *optical switch* can be thought of as a form of optocoupler, as it contains an LED opposite a sensor. However, the LED and the sensor are separated by an open slot, to allow a thin moving object to pass through, interrupting the light beam as a means of detecting the event. It is categorized as a *sensor* in this encyclopedia, and will be found in Volume 3.

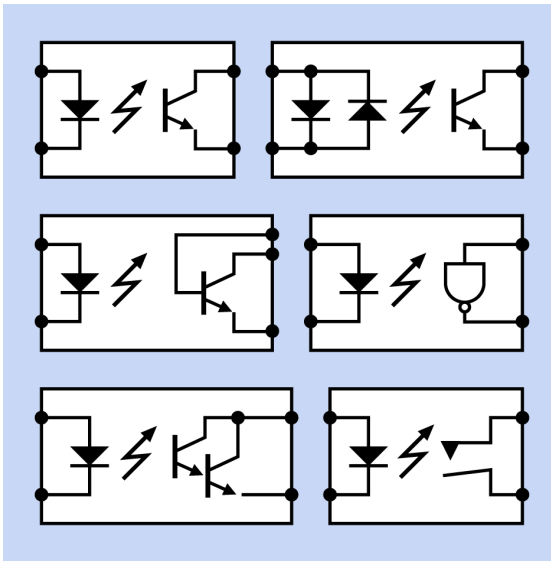


Figure 5-2. Six variants of schematic symbols that may be used to represent an optocoupler. See text for details.

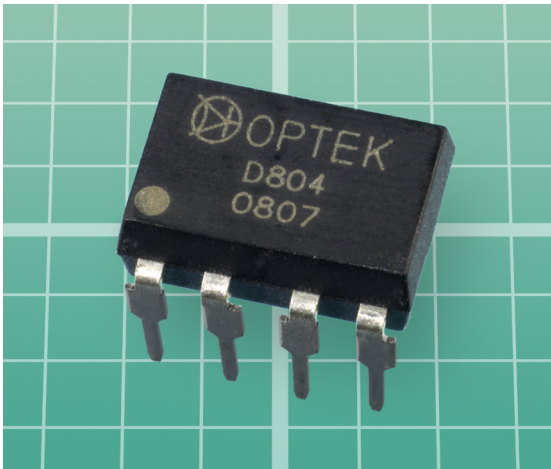


Figure 5-3. An optocoupler in through-hole 8-pin DIP format.

Variants

Internal Sensors

Historically, a photoresistor (often referred to as a *photocell*) was the first type of sensor to be used. It has a more linear response than other sensor types, but its response is much slower. It is still found in audio applications. “Stomp box” pedals

used by guitarists typically contain an optocoupler that employs a photoresistor, and are valued for their linearity and their immunity from the mechanical wear, contamination, and “scratchiness” that builds up over time in a **potentiometer**. Optocouplers also eliminate *ground loops*, which tend to be induced by small differences in ground potential, introducing hum or buzz in audio applications when two or more power supplies are tied together.

The type of optocoupler that contains a photoresistor and is commonly used by musicians was initially trademarked as a *Vactrol*, and that term is still used generically. Vactrols have also been used to provide audio compression in telephone voice networks, and were used in photocopiers and photographic exposure meters, but these applications are now obsolete.

Photoresistors are becoming uncommon because of their cadmium content, which is unlawful in many countries (especially in Europe) because of its environmental toxicity.

A photodiode provides the fastest response time in an optocoupler, limited primarily by the characteristics of the LED that shines light upon it. A *PIN diode* can respond in less than a nanosecond; its acronym is derived from its fabrication from p-type and n-type semiconductor layers with an *intrinsic* layer connecting them. This additional layer can be responsive to light. When the diode is slightly reverse-biased, a photon entering the intrinsic layer can dislodge an electron, enabling current to flow. The reverse bias enlarges the active area and enhances the effect. In this mode, the PIN acts like a photoresistor, appearing to reduce its resistance in response to light.

When the PIN is used in *photovoltaic* mode, no bias is applied, and the component actually generates a small voltage (less than 1VDC), like a solar cell, in response to incoming light. Where an optocoupler uses a MOSFET on its output side, as many as 30 photodiodes may be connected in series to develop the necessary threshold volt-

age to trigger the transistor. This arrangement is common in **solid-state relays**.

A bipolar phototransistor is a slower-speed device but is still usually capable of a 5μs response time or better. Its open collector requires external voltage and a *pull-up resistor* to deliver a positive output so long as the phototransistor is nonconductive. When the LED turns on, the phototransistor sinks current, effectively creating a low output. In this way, the optocoupler functions like an *inverter*, although some variants include a noninverting output.

Basic Optocoupler Types

An optocoupler with *high linearity* will respond more proportionally to variations in current to its LED. *High Speed* optocouplers are used for high-frequency data transfer. *Logic-output* optocouplers have a clean high/low output transition, rather than an *analog output*, which varies with fluctuations in the input. Linearity is of importance only where an optocoupler is being used to transmit an analog signal with some fidelity. Some logic-output optocouplers provide the function of a Schmitt trigger on their output side.

While optocouplers are available in various package formats, the DIP style with six or eight pins remains popular, providing sufficient physical space for the LED, the sensor, and a light channel, while providing good electrical isolation.

Variants may have two or four optocouplers combined in one package. A *bidirectional* optocoupler may consist of two optocouplers in parallel, inverted with respect to each other.

Values

In a datasheet, the characteristics of primary importance in an optocoupler are:

- CTR is the Current Transfer Ratio, the ratio of maximum output current to input current, expressed as a percentage. With a bipolar phototransistor output, 20% is a typical minimum CTR. With a photodarlington output,

the CTR may be 1,000% but the bandwidth is much lower—the response time may be measured in microseconds rather than nanoseconds. Optocouplers with a photodiode output have a very low CTR, and their output is in microamps. However, they provide the most linear response.

- $V_{CE(MAX)}$ is the maximum collector-emitter voltage difference (in an optocoupler with a bipolar phototransistor output). Values from 20 to 80 volts are common.
- V_{ISO} is the maximum potential difference, in VDC, between the two sides of the optocoupler.
- I_{MAX} is the maximum current the transistor can handle, generally in mA.
- Bandwidth is the maximum transmittable signal frequency, often in the range of 20kHz to 500kHz.

The LED in an optocoupler typically requires 5mA at a forward voltage of 1.5V to 1.6V.

The maximum collector current on the output side of an optocoupler is unlikely to be higher than 200mA. For higher output currents, a solid-state relay should be considered. It provides photo-isolation on the same basis as an optocoupler, but high-current versions tend to be considerably more expensive.

How to Use It

The primary purpose of an optocoupler is to provide protection against excessive voltage—from transients, incompatible power supplies, or equipment with unknown characteristics. If a device is designed to be plugged into a USB port on a computer, for instance, the computer may be isolated via an optocoupler.

A series resistor for the LED is not built into most optocouplers, because the value of the resistor will depend on the input voltage that is used. Care must be taken to determine what the maximum voltage on the input side will be, and a

series resistor should be chosen to reduce current appropriately. Allowance should be made for some degradation in the performance of the LED over time.

For an optocoupler with an open-collector output, a pull-up resistor is necessary in most applications. The voltage from the optocoupler must be matched to the input requirements of other components, and the collector current must remain within the specified limits. Some trial and error in resistor selection may be necessary.

In [Figure 5-4](#), a schematic shows typical component values in a test circuit using a pushbutton as input. The separation of the two power supplies is emphasized by the different color shades used for the positive and negative symbols. Although the input side and the output side of an optocoupler may be used with a common ground, this defeats its purpose in providing complete isolation between the sections of the circuit.

The pinouts for an optocoupler must be checked carefully in the manufacturer's datasheet. While the input for an 8-pin DIP chip is usually applied to pins 2 and 3, the output pin functions are not standardized and will vary depending on the internal configuration of the chip. An optocoupler such as the Optek D804, with an enable function using an internal NAND gate, requires its own power supply.

Where an optocoupler allows an external connection to the base of its internal bipolar output phototransistor, reverse bias applied to this pin will decrease the sensitivity of the optocoupler but can increase its immunity to noise on the input side.

What Can Go Wrong

Overload conditions on the input or the output side of an optocoupler will be the most likely cause of failure.

Age

Because optocouplers are typically rated for only 10 years of average use, the age of a component may cause it to fail.

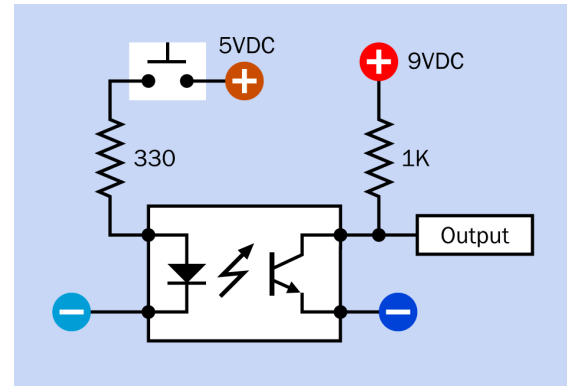


Figure 5-4. Typical values for a series resistor (to protect the LED) and pull-up resistor (to control current and voltage on the output side) in an optocoupler test circuit.

LED Burnout

Because the LED is hidden inside the component, there is no immediate indication of its performance. A meter can be inserted into the circuit on the input side to determine if current is passing through the LED. A meter set to measure volts can be used to discover whether the LED is imposing a normal voltage drop. While significant overload will cause immediate burnout, slightly exceeding the current rating of the LED may have more pernicious consequences, as the LED may not fail until days or weeks have passed without any sign of trouble. The failure of the optocoupler will be unexpected and difficult to determine.

Transistor Burnout

Here again the damage caused by excessive current may be progressive, occurring over a prolonged period. The easiest way to test an optocoupler that may have failed is by removing it from the circuit. A socketed DIP package is preferable for this purpose.

comparator



Although a **comparator** has the same schematic symbol as an **op-amp**, their applications differ and they are described in separate sections of this encyclopedia.

This entry describes an analog comparator. A *digital comparator* is very different, being a logic chip that compares two binary numbers that can be referred to as A and B. Outputs from the chip indicate whether $A > B$ or $A < B$ or $A = B$. The digital comparator does not have an entry in this encyclopedia.

OTHER RELATED COMPONENTS

- **op-amp** (see [Chapter 7](#))

What It Does

A **comparator** is an integrated circuit chip that compares a variable voltage on one input pin with a fixed, reference voltage on a second input pin. Depending which voltage is higher, the output from the comparator will be high or low.

The output will make a clean transition between two fixed values, even if the input is infinitely variable. Thus the comparator can function as an *analog-digital converter*, as shown in [Figure 6-1](#).

Because the output voltage range can be adjusted up or down independently of the input range, a comparator can also function as a *voltage converter*.

Hysteresis

If positive feedback is added through external resistors, *hysteresis* can be introduced. We may imagine a *hysteresis zone* extending above and below the reference voltage level. Small input variations that occur within the zone will be ignored. The comparator only reacts when the input signal emerges above or below the hysteresis zone. When the input signal returns into the hys-

teresis zone, this event also will be ignored. The concept is illustrated graphically in [Figure 6-2](#). A circuit to create hysteresis is shown in [Figure 6-10](#).

How It Works

The schematic symbol for a comparator is shown in [Figure 6-3](#). This seems identical to the symbol for an op-amp, described in [Chapter 7](#), but an op-amp is traditionally a *dual-voltage* device using positive and negative power sources that are equal and opposite, in addition to a zero value midway between the two. Modern comparators mostly use a conventional single voltage, and therefore the negative symbol used in comparator schematics throughout this section of the encyclopedia represents 0 volts. It has the same meaning as the ground symbol found in many schematics elsewhere.

The two inputs to a comparator are described as *inverting* and *noninverting* (for reasons explained later). Confusingly, these are identified with plus and minus symbols inside the triangle that represents the component. These plain black-and-

white symbols have nothing to do with the power supply.

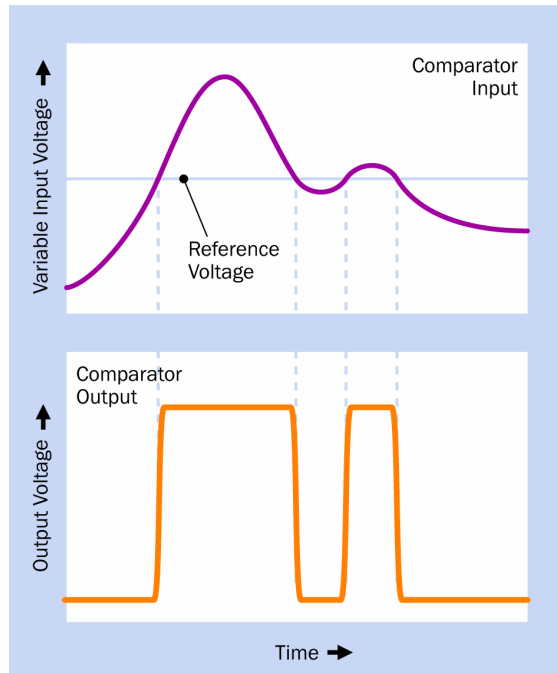


Figure 6-1. The basic behavior of a simple comparator is shown here.

Often, in schematics, the power supply is not shown, because it is assumed to be present. However, all comparators require a power supply in order to function.

The basic internal and external connections used in conjunction with a typical comparator are shown in [Figure 6-4](#).

The potentiometer at top left is often a *trimmer*, to fine-tune a reference voltage. The variable input can come from a sensor or any other device capable of delivering a voltage up to the limit set by V1.

The output is often an *open collector* from an internal bipolar transistor, as shown in the figure.

Note that as many as three different voltages can be used, as indicated by the different colors associated with V1, V2, and V3. However, they must

share a common ground to enable the comparator to make valid comparisons.

When the noninverting input exceeds the voltage of the inverting input, the output transistor goes into its “off” state, and blocks current from an external *pullup resistor*. Because the current from the resistor now has nowhere else to go, it is available to drive other devices attached to the comparator output, and the output appears to be high.

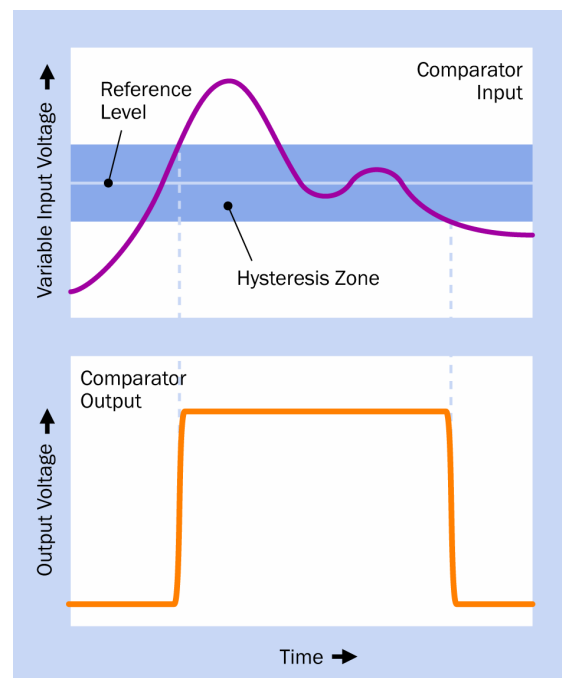


Figure 6-2. The performance of a comparator shown in the previous figure can be modified by the addition of hysteresis. Small variations that occur within the hysteresis zone are ignored.

When the noninverting input falls below the voltage of the inverting input, the transistor becomes conductive, and sinks almost all the current from the pullup resistor, assuming other devices attached to the output have a relatively high impedance. The output from the comparator now appears to be low.

This can be summed up as follows:

- When a variable voltage is applied to the noninverting input, and it rises *above* the reference voltage applied to the *inverting input*, the output transistor turns off, and the comparator delivers a *high* output.
- When a variable voltage is applied to the noninverting input, and it falls *below* the reference voltage applied to the *inverting input*, the output transistor turns on, and the comparator delivers a *low* output.

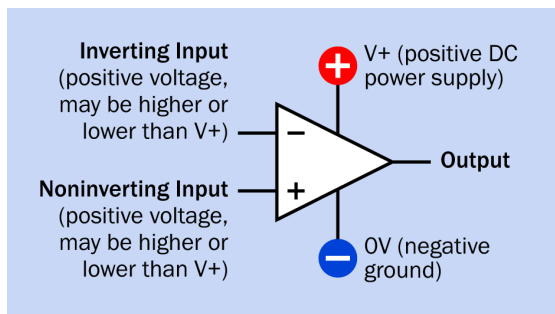


Figure 6-3. The symbol for a comparator is the same as the symbol for an op-amp, even though they often require different types of power supply and their functions are significantly different.

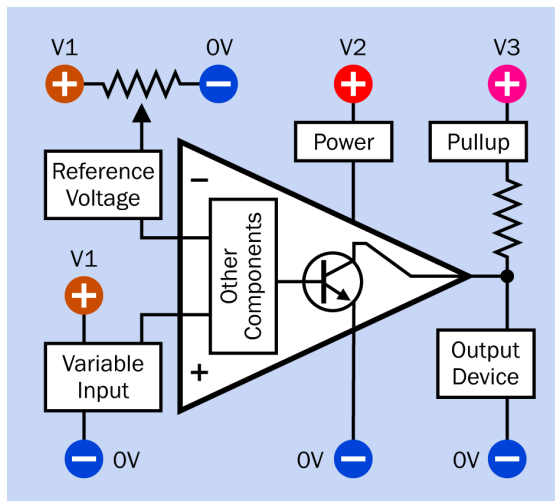


Figure 6-4. Connections to a comparator, and their functions.

If the reference voltage and the variable voltage are swapped between the input pins, the behavior of the comparator is reversed. This relationship is illustrated in **Figure 6-5**. When a voltage transition is applied to the *inverting input*, the transition is *inverted* at the output.

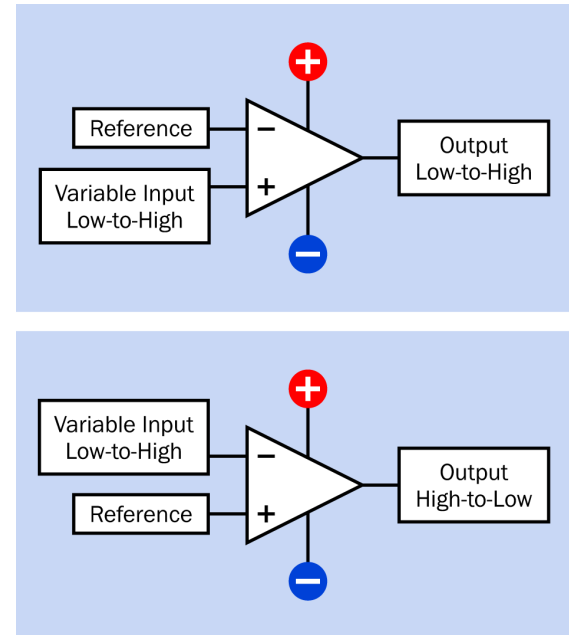


Figure 6-5. Depending which input pin is used for the reference voltage, and which input carries a variable voltage, the comparator output either follows the variable voltage or inverts it.

Placement of the plus and minus signs inside the comparator symbol may vary. Most often, the minus sign is above the plus sign, as shown in all the schematics here. Sometimes, however, for convenience in drawing a schematic, the plus sign may be shown above the minus sign. Regardless of their placement, the plus sign always identifies the noninverting input, and the minus sign always identifies the inverting input. To avoid misinterpretations, schematics should be inspected carefully.

Where a power supply for the comparator is shown, the positive side is always attached to the upper edge of the symbol, while 0V ground is always attached to the lower edge.

Differences from an Op-Amp

Saturation versus linearity

The output of a comparator is optimized for saturation (high or low, without intermediate levels, using positive feedback). The output of an op-amp is optimized for linearity (faithful reproduction of nuances in the input, using negative feedback).

Output mode

The majority of comparators have open-collector outputs (or open-drain outputs in CMOS devices) where the voltage is established by a pullup resistor. This can be adjusted for compatibility with other components, especially 5VDC logic. Only a minority have push-pull amplifier outputs that require no pullup resistor. By comparison, among op-amps, a push-pull output that functions as a voltage source is the traditional default.

Faster response

A comparator responds more quickly than an op-amp to changes in input voltage, if the op-amp is used in the role of a comparator. The comparator is primarily a switching device, not an amplifier.

Hysteresis

This is generally desirable in a comparator, for reasons already explained, and some components are designed with hysteresis built in. This feature is undesirable in an op-amp, as it degrades sensitivity.

Open-loop operation

(i.e., without feedback) this can be used with a comparator. An op-amp is intended for use in closed-loop circuits (i.e., with feedback), and manufacturers will not specify its performance in an open loop.

As previously noted, a comparator usually requires a single-voltage power supply, while an op-amp often requires a dual-voltage power supply.

Variants

Where a comparator uses a MOSFET output transistor, it may have an open-drain output, which requires a pullup resistor, as with an open-collector output.

Some comparators have a *push-pull* output, capable of supplying output current (usually a small amount). In these instances, no pullup resistor is necessary or desirable. The output voltage range will be closest to *rail-to-rail* values (i.e., the range of the power supply) where MOSFETs are used for the output, as MOSFETs impose a smaller voltage drop than bipolar transistors.

The advantage of an open collector (or open drain) relative to a push-pull output is that it allows the output voltage to be set independently of the power supply voltage. Another advantage is that multiple outputs can be connected in parallel, as in a *window comparator* circuit (described below).

Some comparators incorporate a reference voltage on the chip, based on the power supply to the chip. In this case, a separate reference voltage does not have to be supplied, and the component will draw less current.

Many chips are available containing two or more comparators. This is often expressed as the number of *channels* in the component. A *dual comparator* typically allows two different voltage sources for the outputs of the comparators. They will share the same 0V ground, however. Chips such as the LM139 and LM339 contain four comparators, and are available in through-hole or surface-mount formats. They have become a generic choice, costing less than \$1 apiece.

An LM339 comparator chip is shown in [Figure 6-6](#). This is a quad chip, meaning that it contains four comparators. They share a common power supply. The chip is TTL and CMOS compatible, is typically powered by 5VDC, but can be driven by up to 36VDC. The input differential voltage range also extends up to 36V.



Figure 6-6. The LM339 quad comparator chip, shown here, was introduced long ago but remains widely used.

Some comparators have an internal *latch* function that is accessed by a dedicated pin. The latch-enable signal forces the comparator to assess its inputs and hold an appropriate output which can then be checked by other components.

Values

In a datasheet, V_{IO} (also referred to as V_{OS}) is the *input offset voltage*. This is a small voltage, in addition to the reference voltage, which the comparator will require to *toggle* its output in either direction, up or down. Figure 6-7 shows this graphically. V_{IO} sets the limit of *resolution* of the comparator, which will not respond unless the input voltage exceeds the reference voltage by this amount. A smaller value for V_{IO} is better than a larger value. Common values for V_{IO} range from 1mV to 15mV. The actual offset voltage tends to vary between one sample of a component and another. V_{IO} is the maximum allowed value for a component.

Because the comparator will not respond until the reference voltage is exceeded by V_{IO} , the output pulse width will be narrower than if the comparator reacted at the point where the variable voltage input was precisely the same as the reference voltage.

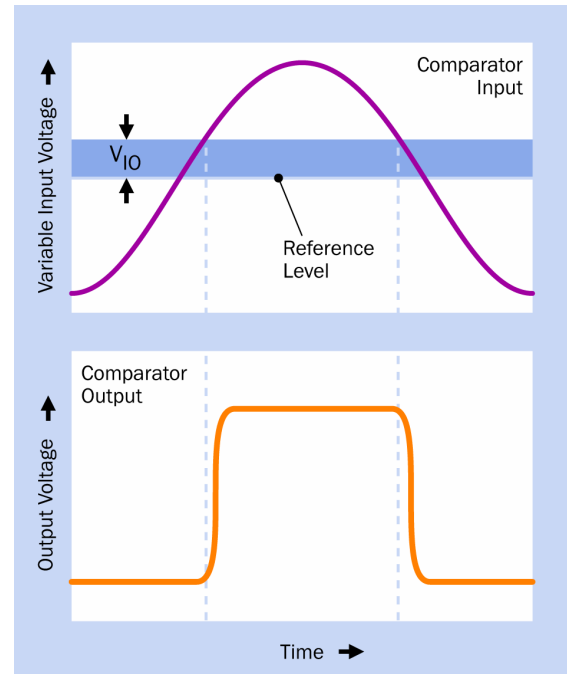


Figure 6-7. The input offset voltage is the very small voltage that a comparator requires, additional to the reference input voltage, before it will toggle its output from low to high or high to low.

V_{TRIP+} and V_{TRIP-} are the *rising* and *falling voltage*s, respectively, that will trip the comparator output where the comparator exhibits some innate hysteresis without an external feedback loop. They are also referred to as *Lower State Transition Voltage* (LSTV) and *Upper State Transition Voltage* (USTV).

V_{HYST} is the *hysteresis range* defined as V_{TRIP+} minus V_{TRIP-} . The relationship is shown graphically in Figure 6-8.

A_{VD} is the *voltage gain* of a comparator, in which the letter “A” can be thought of as meaning “amplification.” The gain is measured as a maximum ratio of output voltage to input voltage. Typically it ranges from 40 to 200.

Supply voltage for modern comparators is often low, as the components are used in surface-mount format for battery-powered devices where low power consumption is a primary con-

cern. Thus, 3VDC is common as a power requirement, and 1.5VDC comparators are available. Still, older chips can use as much as 35VDC.

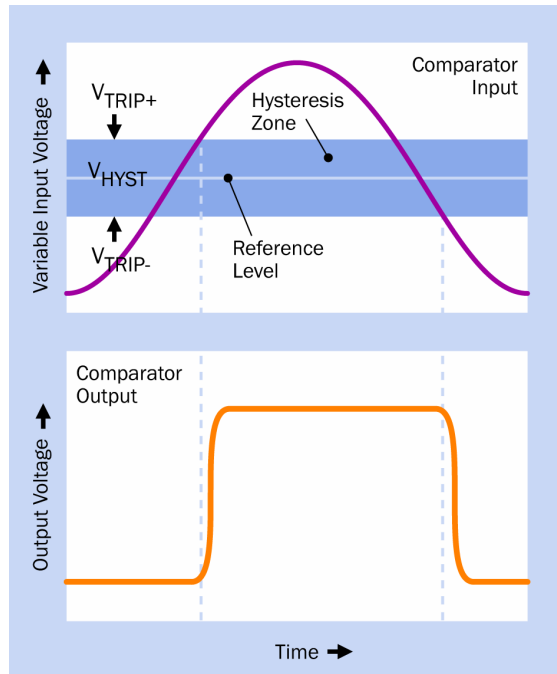


Figure 6-8. The value of V_{TRIP} shows the hysteresis in a comparator—the range of input voltages, relative to the reference voltage, in which it will not respond.

Supply current can range from 7mA down to below 1 μ A.

I_{SINK} is the recommended typical or maximum **sink current** that the component will tolerate, if it has an open-collector output. This value should be considered in relation to the power dissipation, P_D .

The **propagation delay** in a comparator is measured from the moment when an input (usually a square wave) reaches the triggering value, to the time when the consequent output reaches 50% of its final value.

When a comparator is driving CMOS logic using a 5VDC power supply, a typical value for a pullup resistor is 100K. It does not have to be lower, because CMOS has such a high input impedance.

How to Use It

In [Figure 6-1](#), a hypothetical comparator responds immediately when the input voltage equals the reference voltage value. However, this is an idealized scenario. A magnified view, in [Figure 6-9](#), suggests that the comparator is likely to respond with **jitter** when the input signal is very close to the reference voltage, because of tiny variations in heat, current, and other variables. This jitter will cause significant problems if the comparator is driving a device such as a relay, directly or indirectly.

Hysteresis eliminates this uncertainty around the transition level of the input, by telling the comparator to ignore small irregularities in the input voltage. Hysteresis is also useful in many situations where larger variations in a sensor input should be ignored. In [Figure 6-2](#), for instance, suppose that the input voltage comes from a temperature sensor. The small bump in the right-hand section of the curve is probably unimportant; it could be caused by someone opening a door, or a person's body heat in brief proximity to the sensor. There's no point in responding to every little event of this type. In this application, the larger, longer-term temperature trend is what matters, and significant hysteresis is appropriate.

Also, if a comparator is being used as a thermostat, to switch a heating system on and off, we do not want the comparator to respond as soon as the temperature rises just a small amount. The heating system should run for a while before it elevates the temperature beyond the hysteresis zone.

The usual way to create hysteresis is with **positive feedback**. In [Figure 6-10](#), a connection from the output of the comparator runs back through a 1M potentiometer to the variable (noninverting) input. The effect that this has is to reinforce the input voltage with the output voltage, as soon as the comparator input goes high. Now the input can diminish slightly without switching off the comparator. But if the input declines significant-

ly, even the feedback from the output voltage won't be sufficient to maintain the variable input at a higher level than the reference voltage. (Remember, the "high" output voltage from the comparator is a fixed value; it does not change in proportion with the input voltage.) Consequently, the output toggles to low. Now the variable input is deprived of help from the comparator output, so it will be low enough that it has to rise considerably to toggle the comparator back on again. During that period, once again, small variations will be ignored.

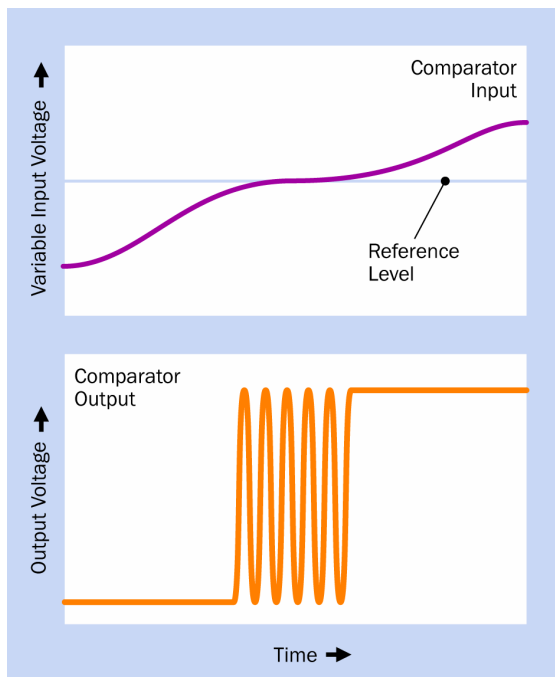


Figure 6-9. In real-world applications, tiny variations where the variable input voltage crosses the reference voltage can induce jitter in the output from a comparator that has no hysteresis.

In the schematic, a **phototransistor** (PT1, at left) is in series with a 3.3K resistor to adjust its voltage output to a suitable range. A 1M potentiometer at upper-left is wired as a voltage divider, so that it can establish a reference level that matches the light level that we wish to detect with the phototransistor.

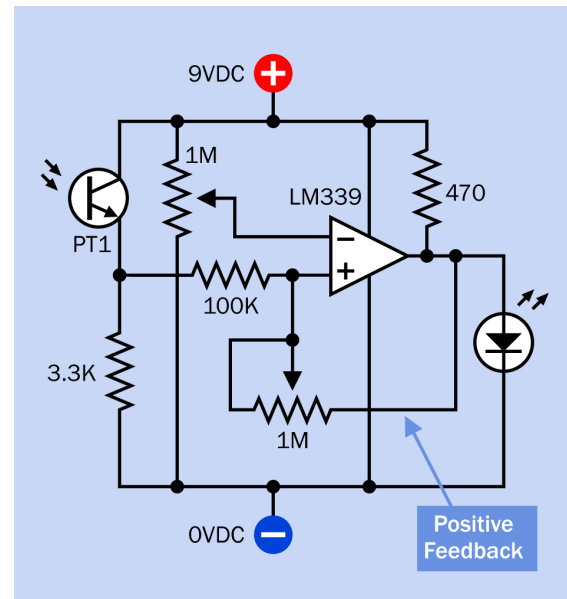


Figure 6-10. A simple circuit to achieve hysteresis with positive feedback to the variable input of a comparator.

The 470Ω resistor is the pullup resistor, which protects the LED from excessive current. The lower 1M resistor adjusts the amount of positive feedback, which determines the width of the hysteresis zone.

Values for components may have to be adjusted depending on the supply voltage, the variable input voltage, and other factors. But the principle will remain the same. Note that in the example shown, all the positive voltage sources are identical. In practice, different voltages could be used, so long as they share a common ground.

AND gate

A set of open-collector comparators can function jointly as an AND gate, when their outputs are tied together with one pullup resistor. So long as all the output transistors are nonconductive, the output will be high. If just one comparator toggles into conductive mode, the output will be low. See [Figure 6-11](#).

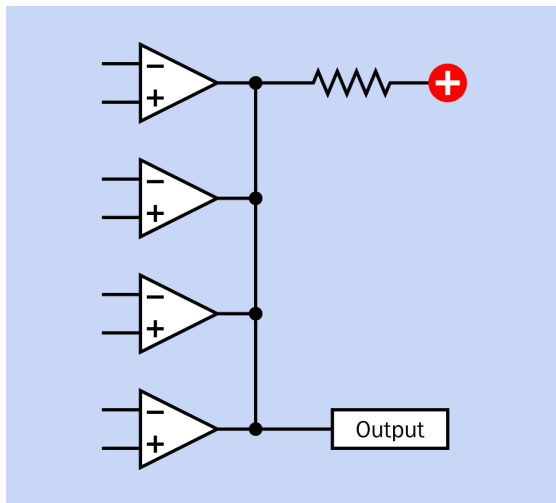


Figure 6-11. If the outputs of multiple open-collector comparators are tied together with a suitable pullup resistor, they will function as an AND gate.

Bistable Multivibrator

If positive feedback to the noninverting input of the comparator is sufficiently high, a voltage almost at 0V ground will be required to counter the high output from the comparator—after which, a voltage almost equal to the supply voltage will be needed to turn it back on. In other words, the comparator is behaving like a *bistable multivibrator*, or **flip-flop**.

Relaxation Oscillator

A *relaxation oscillator*, which is a form of *astable multivibrator*, can be created using direct positive feedback in combination with delayed negative feedback. In **Figure 6-12**, positive feedback goes to the noninverting input, as before, but negative feedback also passes through a 220K resistor to the inverting input of the comparator. A 0.47μF capacitor initially holds the inverting input low, while the capacitor charges. Gradually the capacitor reaches and exceeds the charge on the noninverting input, so the output from the comparator toggles to its low state. This means that its internal transistor is now sinking current, and it discharges the capacitor. Because the noninverting input is being held at a voltage midway between supply and ground by the two 100K re-

sistors forming a voltage divider, eventually the voltage on the inverting input controlled by the capacitor falls below the noninverting voltage, so the cycle begins again.

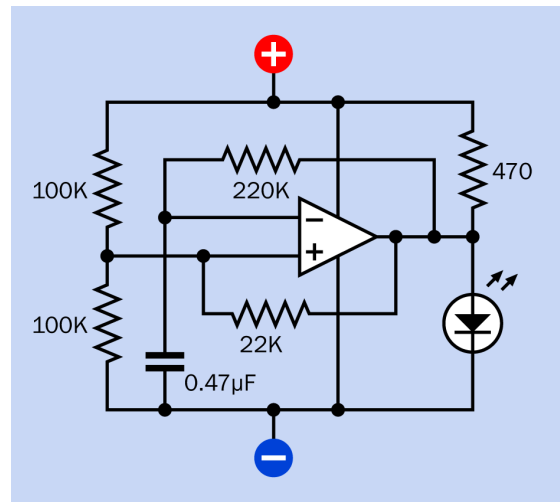


Figure 6-12. A comparator can be used to create a relaxation oscillator.

Level Shifter

Where a comparator is used simply to change the level of an input voltage, it can be referred to as a *level shifter*. An example of a level shifter is shown in **Figure 6-13**, in which a high/low 3VDC logic input is converted to a high/low logic output at 5VDC.

Window Comparator

A *window comparator* is a circuit (not a single component) that will respond to input voltages that deviate outside an acceptable “window” of values. In other words, the circuit responds anytime the variable input is either unacceptably low or unacceptably high.

An example could be an alarm that will sound if a temperature is either too low or too high. In **Figure 6-14**, two comparators are used to create a window comparator circuit, both sharing a variable voltage input from a sensor. A voltage divider is necessary to establish a higher voltage limit at the noninverting input of the upper com-

parator, while a separate voltage divider would establish a lower limit at the inverting input of the lower comparator. If an alarm has an appropriate resistance, it can be used instead of a pull-up resistor. The alarm will sound when the output from either comparator is low, which happens if the inverting input has a higher voltage than the non-inverting input.

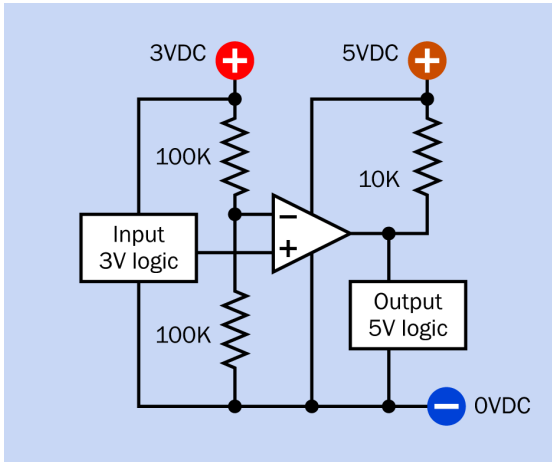


Figure 6-13. A comparator can be used to convert high and low 3V logic inputs into high and low 5V logic outputs.

Other Applications

As previously noted, a comparator can be used as a simple *analog-digital converter*. It has “one bit” accuracy (i.e., its output is either high or low).

A comparator can be used as a *zero point finder* when its variable voltage input is attached to an AC signal. The output from the comparator will be toggled whenever the AC signal passes through zero volts. The output will be a square wave (approximately) instead of a sine wave.

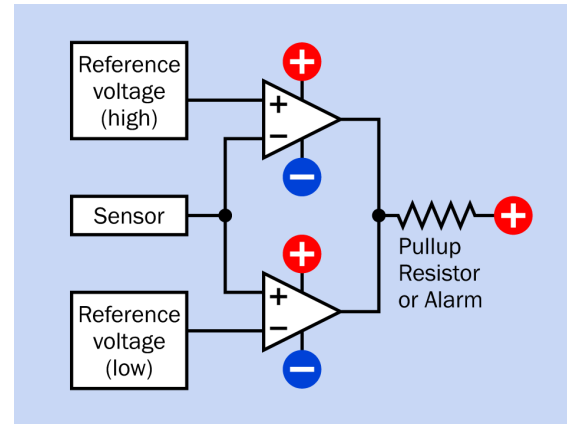


Figure 6-14. A basic, simplified circuit for a window comparator. See text for details.

A *continuous converter* changes its output promptly in response to a change in input. This requires continuous current consumption. Because many applications only need to check the output from a comparator at intervals, power can be saved by using a *clocked* or *latched* comparator.

What Can Go Wrong

Oscillating Output

The high input impedance of a comparator is vulnerable to stray electromagnetic fields. If the conductors leading to and from the comparator are relatively long, the output can couple capacitively with the input during voltage transitions, causing unwanted oscillations.

The commonly recommended solution to this problem is to add 1μF bypass capacitors to the power supply on either side of the comparator. However, some manufacturers recommend alternatives such as introducing just a small amount of hysteresis, or reducing the value of input resistors to below 10K.

If a chip contains multiple comparators, and one of them is unused, one of its input pins should be tied to the positive side of the voltage supply while the other should be tied to 0V ground, to eliminate the possibility of an oscillating output.

Confused Inputs

A comparator will function if its two inputs are swapped accidentally, but its high and low output states will be the inverse of what is expected. Also, if positive feedback is used, transposed inputs can create oscillations. Because the comparator symbol may appear in a schematic with the noninverting input either below or above the inverting input, the inputs are easily transposed by accident.

One way to remember which way the inputs should be connected is to use the mnemonic: “plus, high, positive.” The *plus* input creates a *high* output when the input becomes *more positive* than the reference voltage on the other input. The opposite is less intuitively obvious: the *minus* input creates a *high* output when the input becomes *more negative* than the reference voltage on the other input.

Wrong Chip Type

Different comparators offer different outputs: open collector, open drain, and push-pull. While open collector and open drain function similarly, the pullup resistor value is likely to be different in each case. If a push-pull output is mistakenly connected as if it is open collector or open drain, it will not work correctly, if at all. Different types of comparators must be sorted and stored in clearly labeled bins.

Omitted Pullup Resistor

It is relatively easy to forget to include the pullup resistor on an open-collector output. In this case, when the transistor inside the comparator is in its nonconductive state, the output pin will be

floating, with an indeterminate voltage that will create confusing or random results.

CMOS Issues

As usual when using CMOS chips, it is bad practice to allow unconnected, floating inputs. This is an issue where a chip contains multiple comparators, some of which are not being used. The solution recommended by some manufacturers is to tie one input of an unused comparator to the supply voltage, and the other input of the same comparator to ground.

Erratic Output

If positive feedback is insufficient, the comparator output may show signs of *jitter*. Conversely, if the positive feedback is excessive, the comparator may get stuck in an on state or an off state. Feedback must be chosen carefully.

Swapped Voltages

A comparator is often capable of controlling an output voltage that is much higher than that of its power supply. Because both voltages are applied to different pins on the same chip, mistakes can be made quite easily. The chip is likely to be damaged if the voltages are swapped accidentally between the relevant pins.

Heat-Dependent Hysteresis

Remember that the voltages at which the comparator turns on and off will vary slightly with the temperature of the component. This *drift* should be tested by running the comparator at higher temperatures.

op-amp

7

Although a comparator has the same schematic symbol as an **op-amp**, their applications differ and they are described in separate sections of this encyclopedia.

The unabbreviated name for an op-amp is an *operational amplifier*, but this term is seldom used.

OTHER RELATED COMPONENTS

- **comparator** (see [Chapter 6](#))

What It Does

An op-amp is an *operational amplifier* consisting of multiple transistors packaged in an integrated circuit chip. It senses the fluctuating voltage difference between two inputs, uses power from an external supply to amplify that difference, and uses *negative feedback* to ensure that the output is an accurate replica of the input. Its amplification can be adjusted by changing the values of two external resistors.

Op-amps were developed originally using vacuum tubes, for use in analog computers, before the era of digital computing. Their implementation in integrated circuits dates from the late 1960s, when chips such as the LM741 were introduced (lower-noise versions of it still being widely used today). Multiple op-amps in a single package were introduced in the 1970s.

An LM741 is shown in [Figure 7-1](#). Inside the 8-pin, DIP package is a single op-amp.

How It Works

In alternating current, voltages deviate above and below a zero value, which is sometimes referred to as the *neutral* value. This occurs in do-

mestic power supplies and in audio signals, to name two very common examples. A *voltage amplifier* multiplies the positive and negative voltage excursions, using an external power source to achieve this. Most op-amps are voltage amplifiers.

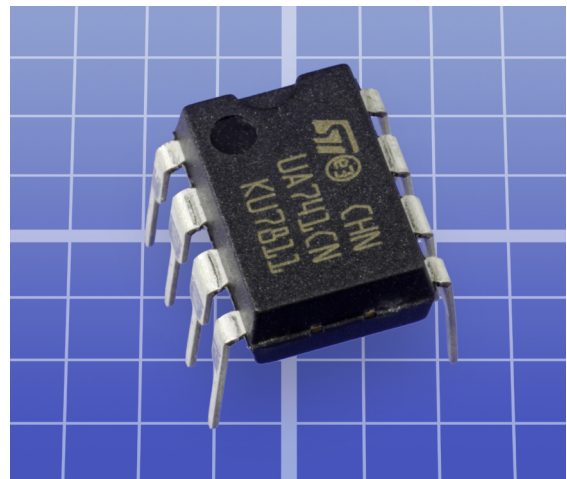


Figure 7-1. The LM741, shown here, is still one of the most widely used op-amps.

An ideal amplifier maintains a *linear relationship* between its input and its output, meaning that the output voltage values are a constant multiple of the input voltages over a wide range. This is

illustrated in [Figure 7-2](#), where the lower curve is a duplicate of the upper curve, the only difference being that its amplitude is multiplied by a fixed amount (usually much greater than shown here). The ratio is properly known as the [gain](#) of the amplifier, usually represented with letter A (for amplification).

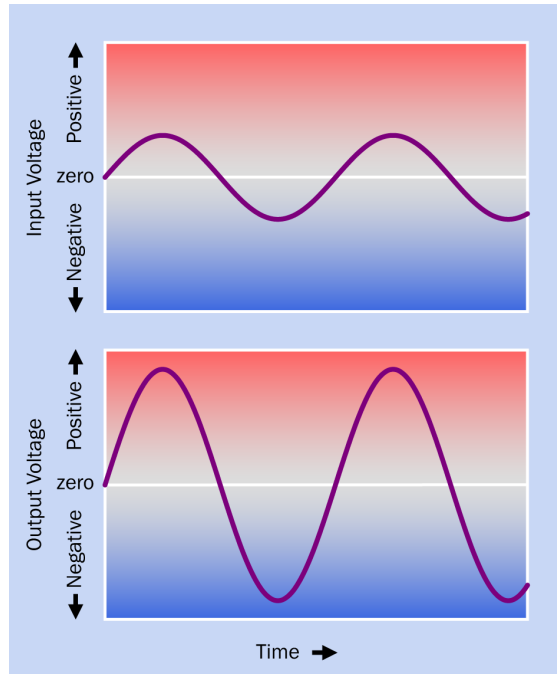


Figure 7-2. In an ideal voltage amplifier, the output voltage will be a duplicate of the fluctuating input voltage, the only difference being that the amplitude of the output is multiplied by a fixed amount. This ratio is known as the [gain](#) of the amplifier.

[Figure 7-3](#) shows the triangular symbol for a generic single-input amplifier (not an op-amp). It may contain any number of components. The triangle almost always points from left to right, with its input on the left and its output on the right, and power attached above and below. This is often a [dual voltage](#) power supply, which is convenient for amplifying a signal that fluctuates above and below 0V. In some schematics, the power supply connections may not be shown, as they are assumed to exist.

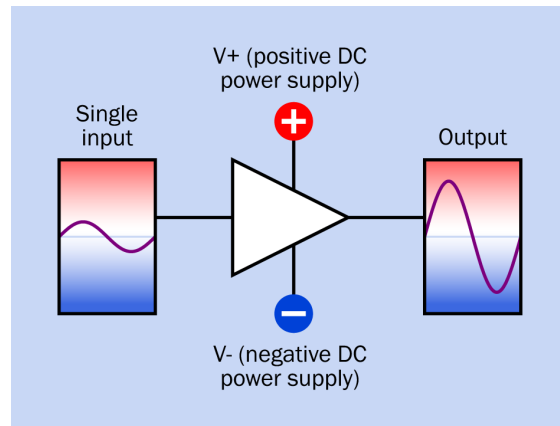


Figure 7-3. The generic symbol for a single-input amplifier (not an op-amp), with the positive side of its power supply being equal and opposite in value to the negative side, and 0V being at the midpoint between them.

- While the blue negative symbol is generally used throughout this encyclopedia to indicate 0V ground, it represents a voltage identified as V– in a dual voltage power supply, being equal in value but opposite in polarity to the positive side of the supply, V+. (Sometimes these voltages are indicated as V– and V+.)

The output from this imaginary generic amplifier is shown in the figure as a linear amplification of the input.

Dual Inputs

An op-amp has two inputs instead of one, and amplifies the voltage difference between them. Its symbol is shown in [Figure 7-4](#). The upper input in this figure is held at 0V, midway between V+ and V–. Because the op-amp has so much gain, an accurate reproduction of its input would create an output exceeding the voltage of the power supply. Because this is not possible, the output tends to become saturated and consequently is [clipped](#) when it reaches its maximum value, as shown in the figure. The thumbnail graphs give only an approximate impression, as they are not drawn to the same scale.

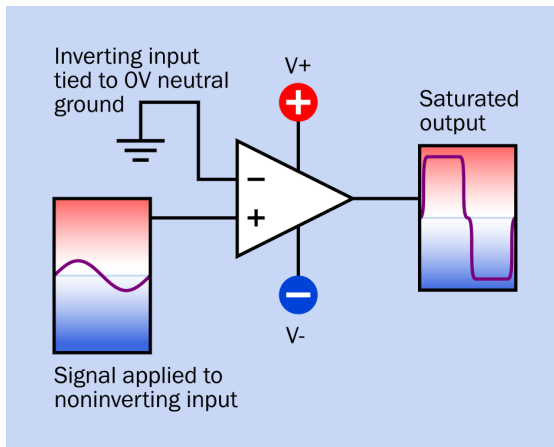


Figure 7-4. An op-amp has so much gain, its output will tend to saturate, producing a square wave regardless of the shape of the input.

The small black plus and minus signs alongside the two inputs to the op-amp have nothing to do with the voltage supplied to the component. The “minus” input is properly referred to as the *inverting input* while the “plus” input is the *noninverting input*, in recognition of their functions.

The inputs are sometimes arranged with the minus above the plus, and sometimes with the plus above the minus. Schematics should be inspected carefully to note which arrangement is being used.

The positive and negative power connections to the op-amp may be omitted, but if shown, they always place V+ at the top, regardless of which way around the inputs are presented.

If a signal is applied to the noninverting input, while the inverting input is held at 0V ground, the op-amp provides an output in which the voltage is not inverted relative to the input.

If the input connections are swapped, so that the inverting input receives the incoming signal while the noninverting input is tied to 0V ground, the output from the op-amp is inverted (the gain remains the same). See [Figure 7-5](#).

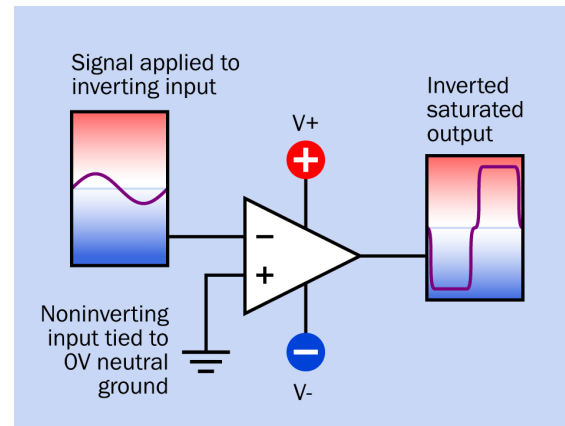


Figure 7-5. When the incoming signal is applied to the inverting input of an op-amp while its noninverting input is held at 0V ground, the output is inverted.

- An op-amp that is being used without any other components to moderate its output is functioning in *open loop* mode.

Negative Feedback

To create an output that is an accurate replica of the input, the op-amp must be brought under control with *negative feedback* to the input signal. This is illustrated in [Figure 7-6](#). A resistor connects the output back to the inverting input, so that the input is automatically reduced to the point where the output is no longer saturated. The values of R1 and R2 will determine the gain of the op-amp, as explained in “[How to Use It](#)” on [page 53](#). The op-amp is now functioning in its intended *closed loop* mode, meaning that the output is being tapped for feedback.

To obtain a linear output that is noninverted, connections are made as shown in [Figure 7-7](#). The resistors form a voltage divider between the output and 0V ground, effectively increasing the comparison value on the inverting input.

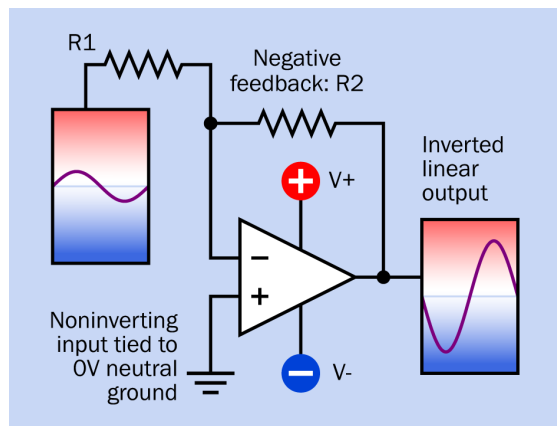


Figure 7-6. A resistor applies negative feedback to the inverting input of an op-amp, and creates a linear output.

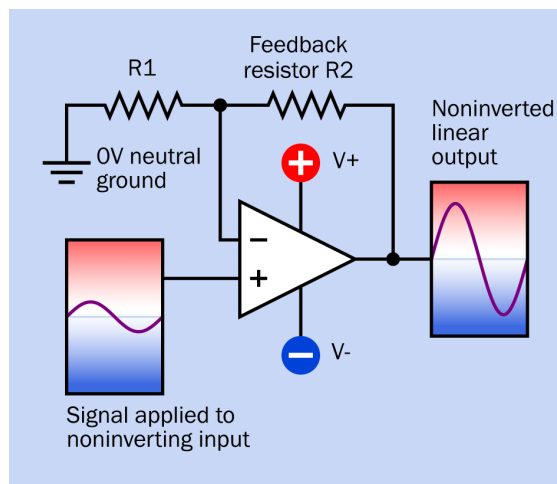


Figure 7-7. Where the incoming signal is applied to the noninverting input, negative feedback is created by using a pair of resistors forming a voltage divider between the output and 0V ground.

- Note that the gain of an op-amp is specific to a particular frequency range of AC signal. This is discussed in [“How to Use It” on page 53](#).

Op-Amps and Comparators

A **comparator** can be regarded as a type of op-amp, and in fact an op-amp can be used as a comparator, comparing a variable DC voltage on one input with a reference voltage on another

input. However, the two types of components have diverged in design to the point where they should be considered separately. The distinction is sufficiently important to have prompted Texas Instruments to issue an Application Report in 2001 titled “Op Amp and Comparators—Don’t Confuse Them!”.

Differences in function are summarized in the previous entry discussing comparators (see [“Differences from an Op-Amp” on page 42](#)).

Variants

Because op-amps are mostly low-current devices, they are widely available in very small surface-mount formats, in addition to the through-hole DIP packages which used to be more common.

Many chips are available containing two or more op-amps. This is often expressed as the number of [channels](#) in the component. A dual chip contains two op-amps, while a quad chip contains four op-amps. Usually all the op-amps in a chip share the same power supply. Bipolar or CMOS transistors may be used.

Because op-amps are widely available in dual and quad packages, it’s quite common for a circuit designer to have one op-amp in a chip “left over.” The designer may be tempted to use that spare unit as a comparator instead of installing an additional chip. To address this situation, some manufacturers offer hybrid op-amp chips containing an additional comparator. The Texas Instruments TLV2303 and TLV2304 are examples.

Values

The op-amps derived from 1970s designs often tolerate a wide range of power-supply voltages. Plus-or-minus 5VDC to plus-or-minus 15VDC is a common range. Modern op-amps are available that run from as little as 1VDC to as much as 1,000VDC.

Op-amps are available for frequencies ranging from 5KHz all the way up to 1GHz.

A “classic” op-amp such as the LM741, which is still widely used, will operate with a power supply ranging from plus-or-minus 5VDC to plus-or-minus 22VDC. Its output is rated for up to 25mA, and its input impedance is at least 2MΩ. The most current it will draw from an input is around 0.5μA.

V_{IO} is the *input offset voltage*. In an ideal component, the output from an op-amp should be 0V when its inputs have a voltage difference of 0V. In practice, the output will be 0V when the inputs differ by the offset voltage. V_{IO} is likely to be no greater than a couple of mV, and negative feedback can compensate for the offset.

V_{ICR} is the *common mode voltage range*. This is the range of input voltages that the op-amp will tolerate. This can never be more than the positive power supply voltage and will often be less, depending on the types of transistors that are used on the input side. If an input voltage goes outside the common mode voltage range, the op-amp will stop functioning.

V_{IDR} is the *input differential voltage range*—the maximum permissible difference between peak positive and peak negative input voltages. This is often expressed as plus-or-minus the power supply voltage, or slightly less. Exceeding the range can have destructive consequences.

I_B is the *input bias current*, averaged over the two inputs. Most op-amps have extremely high input impedance and consequently use very low input currents.

Slew rate at unity gain is the rate of change of the output voltage caused by an instantaneous change on the input side, when the output of the op-amp is connected directly back to the inverting input (during operation in noninverting mode).

How to Use It

In addition to being an amplifier for AC signals, an op-amp can serve as an oscillator, filter, signal conditioner, actuator driver, current source, and voltage source. Many applications require some

understanding of the complexities of mathematics describing alternating current, which are not included in this encyclopedia. Almost all the applications have a common starting point, however, which is to establish and control the gain of the feedback circuit.

Controlling the Gain

A_{VOL} is the *open-loop voltage gain*, defined as the maximum voltage amplification that can be achieved when no feedback is applied from the output to an input. This remains constant until the AC frequency rises to a point known as the *breakover frequency*. If the frequency continues to rise, the maximum gain diminishes quite rapidly, until finally it terminates in 1:1 amplification at the *unity gain frequency*. This transition is shown by the orange line in [Figure 7-8](#). The length of each purple line shows the frequencies which can be tolerated when the op-amp is used in closed-loop mode, and a negative feedback loop limits the gain. For example, where the gain is just 10:1, it can remain constant to just above 10KHz.

Note that both of the scales in this graph are logarithmic.

Calculating Amplification

So long as an op-amp is used within the boundaries of the graph, its voltage amplification can be controlled by choosing appropriate feedback and input resistors. If the op-amp is being used in noninverting mode, and R_1 and R_2 are placed as shown in [Figure 7-7](#), the amplification ratio, A , is found approximately by the formula:

$$A = (\text{approximately}) 1 + (R_2 / R_1)$$

From this it can be seen that when R_1 is very large compared with R_2 , the gain diminishes to near unity. If R_1 becomes infinite and R_2 is zero, the gain is exactly 1:1. This can be achieved by replacing R_2 with a section of wire (theoretically of zero resistance) and omitting R_1 entirely, as in [Figure 7-9](#). In this configuration, the output from the op-amp should be identical with its input.

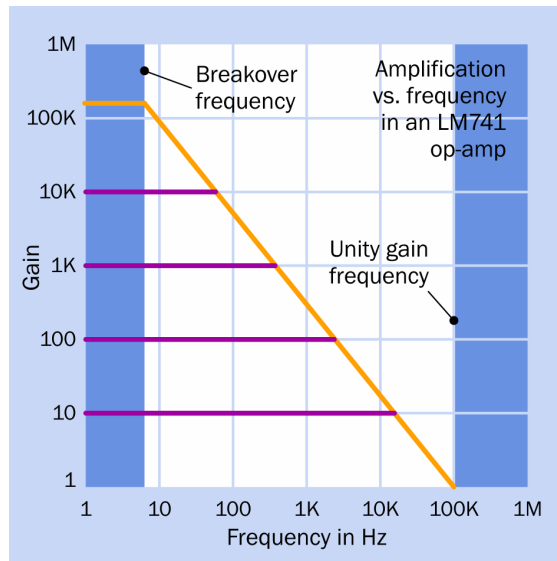


Figure 7-8. Where each horizontal purple line meets the diagonal orange line, this is the maximum frequency that can be used without reduction in the maximum gain of an op-amp.

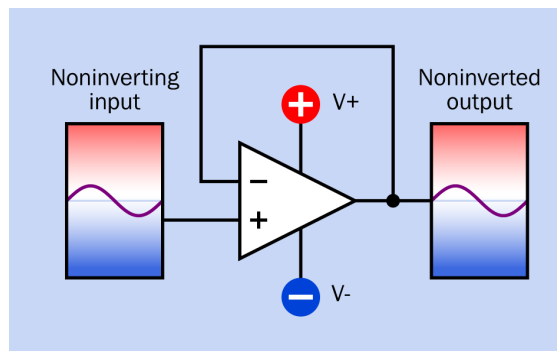


Figure 7-9. While an op-amp is in noninverting mode, if the feedback resistor is replaced with a section of wire and the 0V ground connection is omitted entirely, the gain of the op-amp diminishes theoretically to 1:1.

If the op-amp is being used in inverting mode, and R_1 and R_2 are placed as shown in [Figure 7-6](#), then the voltage amplification ratio, A , is found approximately by the formula:

$$A = (\text{approximately}) -(R_2 / R_1)$$

- Note the minus sign. In inverting mode, gain is expressed as a negative number.

- In a practical circuit, at the expected frequency, the amplification factor established by choice of resistors should be no more than 20.
- An inverting circuit has a relatively low input impedance. For this reason, in most applications, a noninverting circuit is preferred.

Unintentional DC Voltage Amplification

Although the op-amp is intended primarily as an AC signal voltage amplifier, it will also amplify a DC difference between the voltages on its inputs. In the upper section of [Figure 7-10](#), a positive DC offset is inverted and amplified to the point where the output is forced to its negative limit, and the signal is lost, because its fluctuations have been overwhelmed by the positive offset. A coupling capacitor (shown in the lower section of the figure) removes the DC voltage while passing the AC signal. The appropriate capacitor value will depend on the frequency of the signal.

Low-Pass Filter

An op-amp can facilitate a very simple low-pass audio filter, just by adding a capacitor to the basic inverting circuit previously shown in [Figure 7-6](#). The filter schematic is shown in [Figure 7-11](#). Capacitor C_1 is chosen with a value that passes higher audio frequencies and blocks lower audio frequencies. Because the gain of the basic inverting circuit is approximately $-(R_2 / R_1)$, the op-amp functions normally when the impedance of C_1 is blocking the low frequencies, forcing them to pass through R_2 . Higher frequencies, however, are able to bypass R_2 through C_1 , lowering the effective resistance of the feedback section of the circuit, thus reducing its gain. This way, the power of the op-amp is greatly reduced for higher frequencies compared with lower frequencies. A passive RC circuit could achieve the same effect, but would attenuate the signal, while the op-amp circuit boosts part of it.

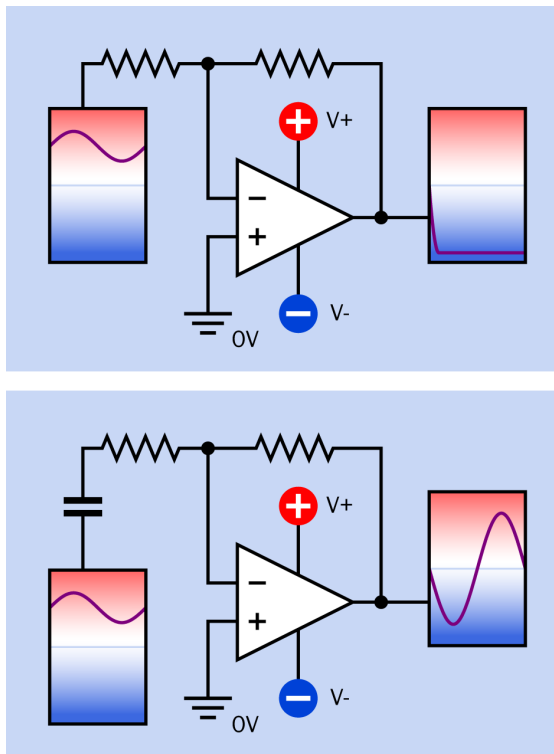


Figure 7-10. The addition of a capacitor at the input of an op-amp is often necessary to prevent any DC voltage offset being amplified. In the upper section of this figure, a DC offset is large enough to force the inverted output to its negative limit, and the signal is completely lost.

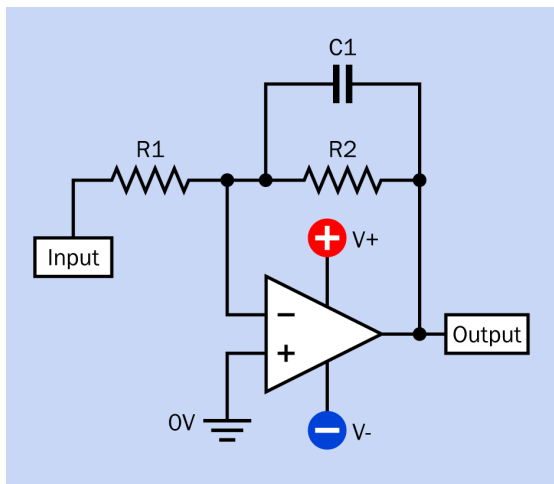


Figure 7-11. A very basic low-pass filter, which works by allowing capacitor C1 to bypass resistor R2 at higher audio frequencies.

High-Pass Filter

A simple high-pass filter can be created by adding a capacitor to the basic noninverting circuit previously shown in [Figure 7-7](#). The filter schematic is shown in [Figure 7-12](#). Once again capacitor C1 is chosen with a value that passes higher audio frequencies and blocks lower audio frequencies. Because the gain of the basic noninverting circuit is approximately $1 + (R2 / R1)$, the op-amp functions normally when the impedance of C1 is blocking the low frequencies, forcing them to pass through R1. Higher frequencies, however, are able to bypass R1 through C1, lowering the effective resistance of that section of the circuit, thus reducing the negative feedback and increasing the gain. This way, the power of the op-amp is increased for higher frequencies compared with lower frequencies. A passive RC circuit could achieve the same effect, but would attenuate the signal, while the op-amp circuit boosts part of it.

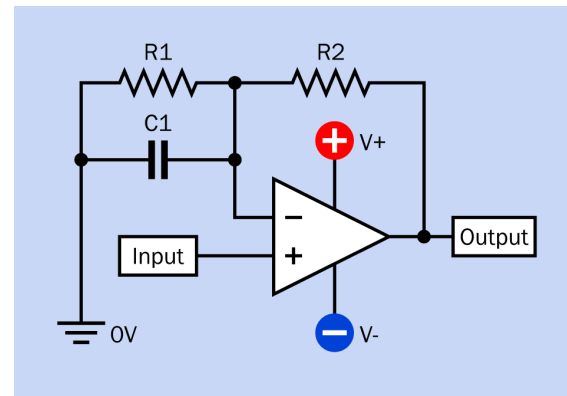


Figure 7-12. A very basic high-pass filter, which works by allowing capacitor C1 to bypass resistor R1 at higher audio frequencies.

Relaxation Oscillator

The schematic in [Figure 7-13](#) is similar to the circuit shown in [Figure 6-12](#) using a comparator. It functions as a *relaxation oscillator*, which is a form of *astable multivibrator*. The lower half of the circuit is a positive feedback loop that reinforces the output while the upper half of the circuit is charging the capacitor. Eventually the charge on the

capacitor exceeds the voltage on the noninverting input of the op-amp, creating negative feedback that exceeds the positive feedback. The capacitor discharges and the cycle repeats. The component values in the figure should generate an output that runs at around 2Hz. Reducing the value of the capacitor will increase the frequency.

Single Power Source

A few op-amps are designed to work from single voltages, but they are a relatively small minority, and will clip the output signal if the input goes negative. Power supplies are readily available that provide multiple voltages such as +15VDC, 0V, and -15VDC. They are ideal for driving an op-amp—but may not be useful for any other components in the circuit. Can an op-amp that is designed for dual voltages be made to run from a single supply, such as 30VDC?

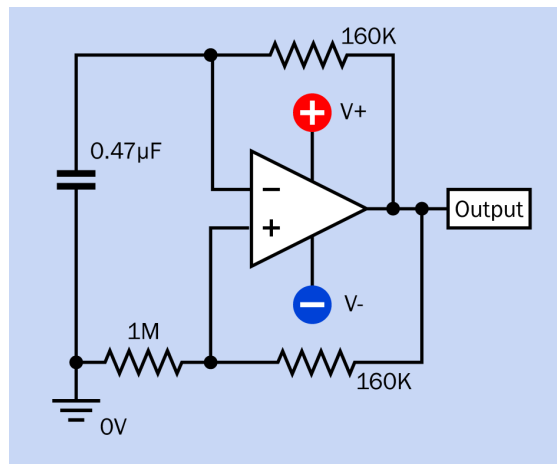


Figure 7-13. A relaxation oscillator.

This is relatively easy to do. The op-amp simply needs a potential difference to power its internal transistors, and 30VDC on the V+ pin with 0VDC on the V- pin will work just as well as +15VDC and -15VDC. However, referring back to [Figure 7-6](#), if the op-amp is used in inverting mode, an intermediate voltage must be supplied to the noninverting input. Likewise, in noninverting mode, an intermediate voltage is necessary for one of the inputs, and must be half-way

between the extremes of the power supply. If the supply is +15VDC and -15VDC, the midpoint is 0V. If the supply is 30VDC and 0V, the midpoint is 15VDC.

Because the inputs of an op-amp have a very high impedance and draw negligible current, the intermediate voltage can be provided with a simple voltage divider, as shown in [Figure 7-14](#), where R3 and R4 should be no greater than 100K each. Their exact values are not important, so long as they are equal.

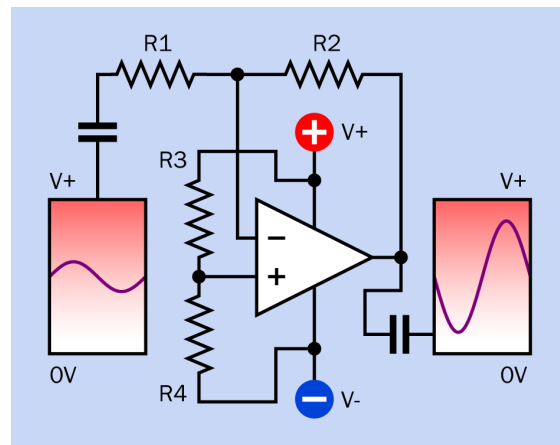


Figure 7-14. A voltage divider, comprised of R3 and R4 in this schematic, can provide a voltage halfway between V+ and negative ground, enabling the op-amp to use just one power supply instead of two.

A coupling capacitor should still be used on the input side, as shown, because there is no guarantee that the input signal will be centered precisely on 15V, and any offset will be amplified, potentially causing clipping of the signal. For similar reasons, a coupling capacitor is also added on the output side.

Offset Null Adjustment

Some op-amps provide two pins for *offset null adjustment*, which is a setup process to ensure that identical voltage on the two inputs will produce a null output. This is a way of compensating for any internal inconsistencies introduced during the manufacturing process.

To perform offset null adjustment, both input pins are connected directly to 0V ground, and the ends of a trimmer potentiometer (typically, 10K) are connected with the offset null pins, while the wiper of the potentiometer is centered and then connected with the negative power supply. The probes of a meter that is set to measure DC volts are placed between the output of the op-amp and 0V ground. The potentiometer is then adjusted until the meter shows a reading of 0VDC. A schematic is shown in [Figure 7-15](#).

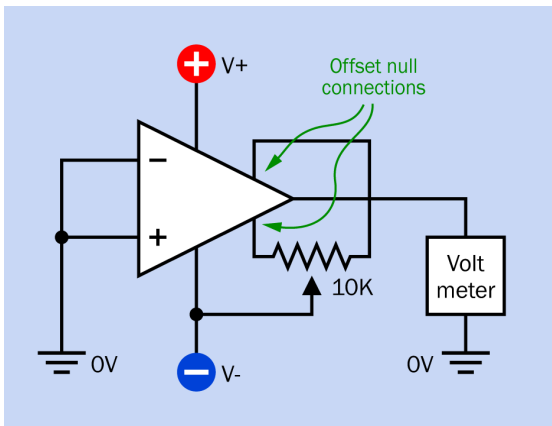


Figure 7-15. Connections for making an offset null adjustment to an op-amp that allows this procedure.

What Can Go Wrong

Power Supply Problems

Op-amps are especially vulnerable to reversed polarity in a power supply. If there is even a remote possibility of this occurring, a diode in series with one side of the supply can provide protection.

A more realistic concern is the destructive consequence of an input signal that exceeds the power supply voltage(s) of the op-amp. Even if the input is within the acceptable range, it can still cause permanent damage if it is applied before the op-amp powers up.

Bad Connection of Unused Sections

Multiple op-amps are often combined in a single package. If some of these “sections” remain unused, they will still receive power from the shared supply, and will attempt to function. If the inputs are left unconnected, they will pick up small stray voltages by capacitance or induction, and in the absence of negative feedback, the op-amp will create unpredictable outputs, consuming power and possibly interacting with other sections of the same chip. [Figure 7-16](#) shows three incorrect options for addressing this problem, and one recommended option (derived from Texas Instruments Application Report SLOA067).

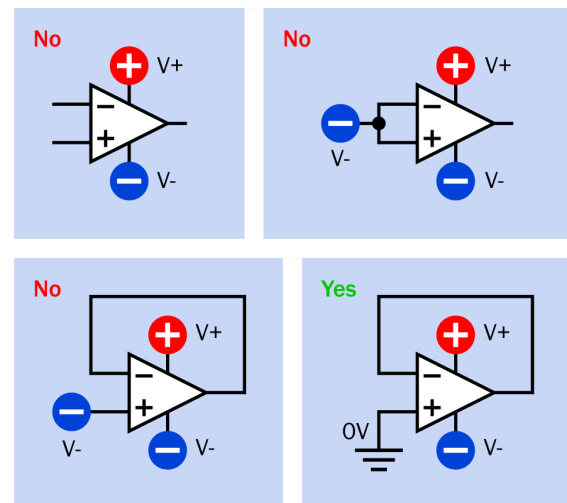


Figure 7-16. When multiple op-amps share a chip, one that is unused will still receive power from the shared supply. Its inputs must not be allowed to float, and must be connected to minimize activity and power consumption. Three common errors are shown here, with one recommended option. Note the distinction between 0V ground (0V) and negative power (V-).

Oscillating Output

The inputs of an op-amp are vulnerable to stray electromagnetic fields. If the conductors leading to and from the op-amp are relatively long, the output can couple capacitively with the input during voltage transitions, causing unwanted oscillations.

The commonly recommended solution to this problem is to add a $1\mu\text{F}$ bypass capacitor between the power supply and 0V ground. However, some manufacturers recommend alternatives such as introducing a very small amount of hysteresis, or reducing the value of input resistors to below 10K.

Confused Inputs

A schematic may show an op-amp with the non-inverting input above the inverting input, or vice

versa. The only indication of this will be the plus and minus signs inside the chip, which can be extremely small and easily overlooked. For convenience in drawing a diagram, two op-amps in the same circuit may have their inputs shown in opposite configurations. Special care must be taken to verify that the inverting and noninverting input pins on a chip are correctly assigned.



digital potentiometer

A **digital potentiometer** is also known as a *digitally adjustable potentiometer*, a *digitally controlled potentiometer*, a *digitally programmed potentiometer* (with acronym *DPP*), a *digipot*, or a *digipot*. The terms are functionally interchangeable. Because the abbreviation *pot* is often used to describe an analog potentiometer, some people refer to digital potentiometers colloquially as *digital pots*. In printed documentation, the letters in *pot* may be capitalized. Because it is an abbreviation, not an acronym, it is not capitalized here.

Because this component enables digital control of a variable voltage, it is a *mixed signal device*. It is classified here as an analog chip because it primarily emulates the function of an analog device. It may be thought of as a form of *digital-analog converter*, although this encyclopedia does not have a section devoted to that type of component or to *analog-digital converters*, as their application is relatively specialized.

OTHER RELATED COMPONENTS

- **potentiometer** (see Volume 1)

What It Does

This component is an integrated circuit chip that emulates the function of an analog **potentiometer**. It is often described as being *programmable*, meaning that its internal resistance can be changed via a control input.

Digital potentiometers are particularly suited for use in conjunction with a *microcontroller*, which can control the internal resistance of the component. Possible applications include adjustment of the pulse width of an oscillator or multistable multivibrator (e.g., using the Control pin of a 555 **timer** chip); adjustment of the gain in an **op-amp**; specification of voltage delivered by a *voltage regulator*; and adjustment of a *band-pass filter*.

A digital potentiometer in combination with a microcontroller may also be used in conjunction with a pair of external buttons or a rotational en-

coder, to adjust the gain of an audio amplifier and for similar applications.

Advantages

A digital potentiometer offers significant advantages over an analog potentiometer:

- **Reliability.** The digital component may be rated for as many as a million cycles (each storing the wiper position in an internal memory location). An analog component may be capable of just a few thousand adjustment cycles.
- **Digital interface.**
- **Elimination of long signal paths or cable runs.** The digital potentiometer can be placed close to other chips, whereas an analog potentiometer often has to be some distance away to enable control by the end user. Reduction in the length of signal paths can reduce capacitive effects, while elimina-

tion of cable runs will reduce manufacturing costs.

- Reduction in size and weight compared with a manual potentiometer.

A digital potentiometer also has some disadvantages:

- Its internal resistance is somewhat affected by temperature.
- It is not usually capable of passing significant current. Few chips can sink or source more than 20mA at the output, and 1mA is common. The output is primarily intended for connection with other solid-state components that have high impedance.
- Users may prefer the immediacy and tactile feel of a knob attached to an analog potentiometer, rather than a pair of buttons or a rotational encoder.

How It Works

A digital potentiometer changes the point at which a connection is made along a *ladder* of many fixed resistors connected in series inside the chip. Each end of the ladder, and each intersection between two adjacent resistors, is known as a *tap*. The pin that can connect with any of the taps is referred to as the *wiper*, because it emulates the function of a wiper in an analog potentiometer. In reality, a digital potentiometer does not contain a wiper or any other moving parts.

A fully featured digital potentiometer allows access to each end of the ladder through two pins that are often labeled “high” and “low,” even though they are functionally interchangeable (except in the case of a component that simulates a logarithmic taper, as described later). The “low” end of the ladder is sometimes numbered 0. In this case, if there are n resistors, the “high” end of the ladder will be numbered n . Alternatively, if the “low” end of the ladder is numbered 1, and there are n resistors, the “high” end will be

numbered $n+1$. This principle is illustrated in Figure 8-1.

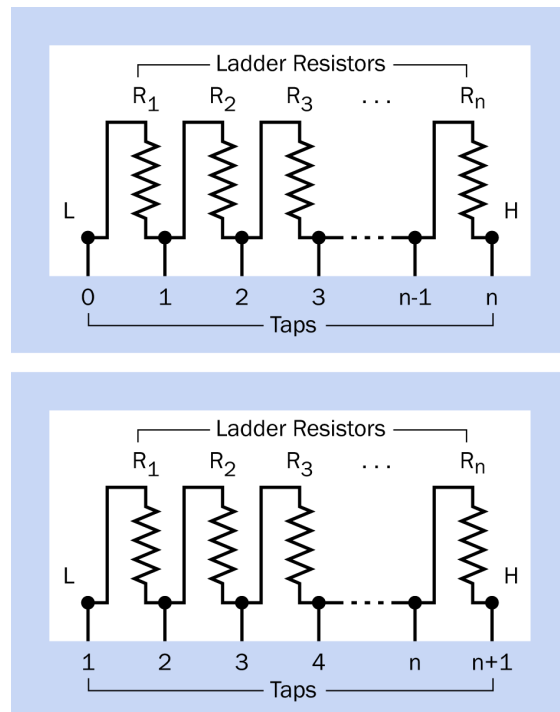


Figure 8-1. Available wiper connections to a resistor ladder inside a digital potentiometer, showing two numbering systems that may be used.

The “low” pin on a digital potentiometer may be identified as L, or A, or R_L , or P_A in a datasheet, while the “high” pin may be identified as H, or B, or R_H , or P_B and the pin that accesses the wiper is typically identified as W, or R_W , or P_W . Letters L, H, and W are used below. Although the L and H pins are functionally interchangeable, their labels are useful to identify which direction the W connection will move in response to an external signal.

Digital potentiometers are available with as few as 4 or as many as 1,024 taps, but common values are 32, 64, 128, or 256 taps, with 256 being the most common.

No specific schematic symbol represents a digital potentiometer. Often the component is shown as an analog potentiometer symbol inside a box that has a part number, as suggested in

Figure 8-2. Control pins and the power supply may be omitted if the schematic is just intended to show logical connections. Alternatively, if the digital potentiometer is depicted in a schematic where it is connected with other components such as a microcontroller, multiple pins and functions may be included, as shown in [Figure 8-3](#). The pins additional to L, H, and W are explained below.

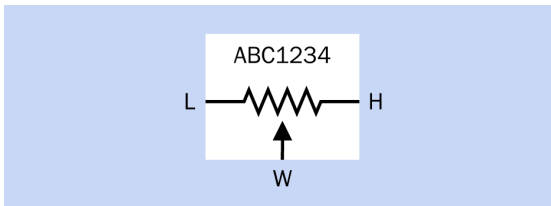


Figure 8-2. There is no single specific symbol to represent a digital potentiometer. It may be shown using an analog potentiometer symbol in a box with a part number, as suggested here, where power connections and additional pins are omitted for clarity.

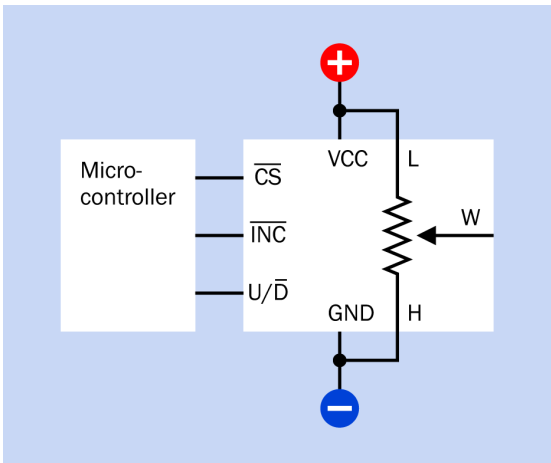


Figure 8-3. If a digital potentiometer appears in a schematic where it is connected with other components such as a microcontroller, additional pins and functions may be indicated. This generic representation of a digital potentiometer shows some of the functions that can be included.

Variants

A *dual* digital potentiometer contains two complete units, while a *quad* contains four. Triples

exist but are relatively uncommon. A few chips contain six potentiometers. Multiple digital potentiometers in a chip can be used as the digital equivalent of *ganged* analog potentiometers, for simultaneous synchronized adjustment of multiple inputs in an audio system (two channels in a stereo amplifier, or more in a surround-sound system).

The pinouts of a sophisticated quad digital potentiometer chip are shown in [Figure 8-4](#). Other quad chips have different pinouts and capabilities; there is no standardized format as there is with digital logic chips. In this example, the high/low states of Address 0 and Address 1 select one of the four internal resistor ladders, numbered 0 through 3. The Chip Select pin makes the whole chip either active or inactive. The Write Protect pin disables writing to the internal wiper memory. The Serial Clock pin inputs a reference pulse stream to which the serial input data must be synchronized. The Hold pin pauses the chip while data is being transmitted, allowing the data transmission to be resumed subsequently. The NC pins have no connection.

Volatile and Nonvolatile Memory

Any type of digital potentiometer requires memory to store its current wiper position, and this memory may be volatile or nonvolatile. Nonvolatile memory may be indicated in a datasheet by the term *NV*.

A digital potentiometer with volatile memory will typically reset its wiper to a center-tap position if power is disconnected and then restored. A digital potentiometer with nonvolatile memory will usually restore the most recently used wiper position, provided the chip is fully powered down and then fully powered up without glitches in the supply. If a microcontroller is being used to control the digital potentiometer, it can store the most recent resistance value in its own nonvolatile memory, in which case the type of memory in the potentiometer becomes irrelevant.

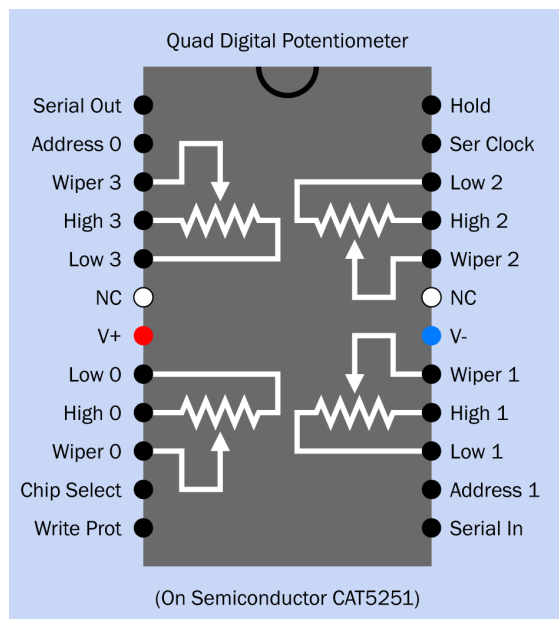


Figure 8-4. Pinouts of a sophisticated quad digital potentiometer chip. Other chips will have different pinouts and capabilities. This example is available in surface-mount formats only. See text for details.

Taper

Digital potentiometers are available with *linear taper* or *logarithmic taper*. In the former, each resistor in the ladder has the same value. In the latter, values are chosen so that the cumulative resistance between the wiper and the L end of the ladder increases geometrically as the wiper steps toward the H end of the ladder. This is useful in audio applications where sound intensity that increases exponentially may seem to increase linearly when perceived by the human ear.

A microcontroller can emulate logarithmic steps by skipping some taps in the ladder in a digital potentiometer, but this will result in fewer increments and lower precision.

Data Transfer

Digital potentiometers are mostly designed to use one of three serial protocols:

- **SPI.** This acronym is derived from *serial peripheral interface*, a term trademarked by Motorola but now used generically. The stan-

dard is adapted in various radically different ways among digital potentiometers.

- **I²C.** More correctly printed as *I²C* and properly pronounced “I squared C,” this acronym is derived from the term *inter-integrated circuit*. Developed by Philips in the 1990s, it is a relatively slow-speed bus-communication protocol (up to 400kbps or 1Mbps in its basic form). It is built into some microcontrollers. The standard is more uniformly and rigorously defined than SPI.
- **Up/down**, also sometimes known as *push-button* or *increment/decrement* protocol.

Both SPI and I²C are supported by many microcontrollers, including the Atmel AVR at the heart of the Arduino.

These three systems for controlling a digital potentiometer are described in more detail in the following sections.

SPI

This is the most widely used protocol, but when reading datasheets, a lot of care must be taken to determine how it varies in each case.

The Microchip 4131-503, shown in [Figure 8-5](#), uses SPI protocol. It contains 128 resistors and can be powered by 1.8VDC to 5.5VDC.

The one feature that all versions of SPI have in common is that a series of high/low pulses is interpreted by the chip as a set of bits whose value defines a tap point in the resistor ladder. In computer terminology, every tap point has an *address*. The incoming bits define the address, after which the status of an additional input pin can tell the chip to move the wiper to that location.

Typically, there will be a chip select pin, identified as CS; a serial data input pin, identified as SDA, SI, DIN, or a similar acronym; and a serial clock pin, identified as SCL, SCLK, or SCK, which must receive a stream of pulses to which the high/low data input pulses must be synchronized. In addition, the SPI protocol allows bidirectional (duplex) serial communication. Only a minority of

digital potentiometers make use of this capability, but where it exists, a serial data output pin may be labeled SDO. Alternatively, one pin may be multiplexed to enable both input and output, in which case it may be labeled SDI/SDO.

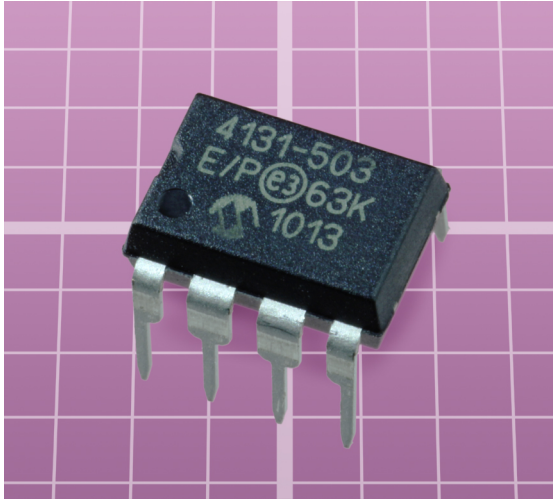


Figure 8-5. This digital potentiometer uses SPI protocol. See text for details.

If a pin is active-low, a bar (a horizontal line) will be printed above its acronym.

The most common type of digital potentiometer has 255 resistors and therefore 256 tap points, which have addresses numbered 0 through 255, each of which can be specified by a sequence of eight data bits constituting one byte. However, a different coding system will be applied in chips that have a different number of taps. In a 32-tap component, for instance, data is still sent in groups of eight bits, but only the first five bits define a *tap address*, while the remaining three are interpreted as commands to the chip.

Most 256-tap chips use an SPI protocol in which two eight-bit bytes are sent, the first being interpreted by the chip as a command, while the second specifies a tap address. Each manufacturer may use a different set of command codes, and these will vary among chips even from the same manufacturer.

Most commonly, three wires are used for data transmission and control (causing these chips to be described as *3-wire* programmable potentiometers).

CS is usually, but not always, pulled low to activate the digital potentiometer for input. A series of low or high states is then applied to the data-input pin. Each time the clock input changes state (usually on the rising edge of the clock pulse) the state of the data input is copied to a **shift register** inside the chip. After all the bits have been clocked in, CS can change from low to high, causing the contents of the shift register to be copied into a decoder section of the chip. The first bit received becomes the most significant bit in the decoder. The value of the eight bits is decoded, and the chip connects the W pin directly to the corresponding tap along the ladder of 255 internal resistors.

I2C Protocol

The I2C specification is controlled by NXP Semiconductors (formerly Philips), but can be used in commercial products without paying licensing fees. Only two transmission lines are required: one carrying a clock signal, the other allowing bidirectional data transfer synchronized with the clock (although many digital potentiometers use the I2C connection only to receive data). The pins are likely to be identified by the same acronyms as the pins on a chip that uses SPI protocol.

As in SPI, a command byte is followed by a data byte, although the command set differs from that of SPI and will also differ among various I2C chips. Full implementation of I2C allows multiple devices to share a single bus, but this capability may remain unused.

Up/Down Protocol

This simpler, asynchronous protocol does not require a clock input. The chip will respond to data pulses that are received at any speed (up to its maximum speed), and the pulse widths can be inconsistent.

Each pulse moves the wiper connection one step up or down the ladder. While this has the advantage of simplicity, the taps are not addressable, and consequently the wiper cannot skip to any tap without passing through intervening taps incrementally. This is not an inconvenience when the potentiometer controls audio gain, which is a primary application.

In some chips, an increment pin, usually labeled INC, receives pulses while the high or low state of a second pin, usually labeled U/D, determines whether each pulse will step the wiper up the ladder or down the ladder.

In other chips, pulses to an Up pin will step the wiper up the ladder, while pulses to a Down pin will step the wiper down the ladder.

Either of these chip designs can be referred to as a *two-wire* type. If an additional chip-select pin is included (labeled CS on datasheets), this type of digital potentiometer can be referred to as a *three-wire* type. The chip select pin is likely to be active-low, meaning that so long as it has a high state, the chip will ignore incoming signals.

The CAT5114 shown in [Figure 8-6](#) uses an U/D pin. It contains 31 resistors, is available in 8-pin DIP or surface-mount formats, and can be powered by 2.5VDC to 6VDC. Each of its logic inputs draws only 10μA.

In six-pin chips the INC pin is omitted, one of the H, L, or W pins will be omitted, and the U/D pin will function differently. When CS is pulled low, the chip checks the state of the U/D pin. If it is high, the chip goes into increment mode; if it is low, the chip goes into decrement mode. So long as CS remains low, each transition of the U/D pin from low to high will either increment or decrement the wiper position, depending on the mode that was sensed initially. When CS goes high, further transitions on the U/D pin will be ignored until CS goes low again, at which point the procedure repeats.

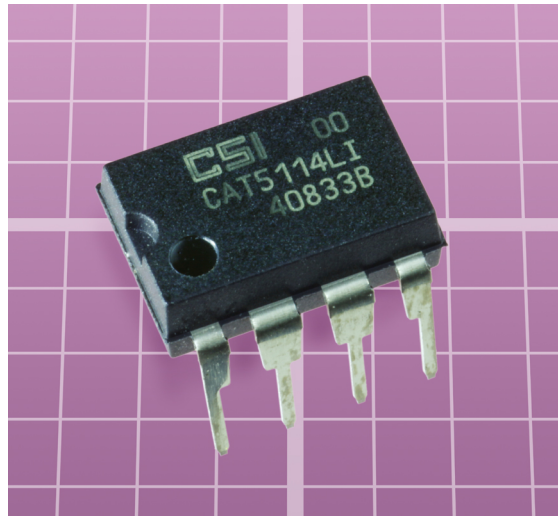


Figure 8-6. This digital potentiometer contains 31 resistors and uses the simplest up/down protocol to step from one tap to the next.

The chip does not provide any feedback regarding the position of its wiper, and consequently a control device such as a microcontroller cannot know the current wiper position. If the chip has nonvolatile memory (as is the case in many up/down digital potentiometers), it will resume its previous wiper location at power-up, but here again a control device will have difficulty determining what that position is. Therefore, in its basic form, an up/down chip is only appropriate for simple tasks, especially in response to up/down pushbuttons.

Other Control Systems

A few digital potentiometers use a parallel interface. Because they are relatively uncommon, they are not included here.

Connections and Modes

Some variants of digital potentiometers minimize the chip size and number of connections by limiting accessibility to the internal resistor ladder. In a chip designed to function in *rheostat mode*, the W pin is eliminated and the chip moves an internal connection point to change the resistance between the H and L pins.

In some variants, the low end of the ladder is permanently, internally connected with ground, and the L pin is omitted. In other variants, one end of the ladder is unconnected inside the chip.

A chip designed to function in *voltage divider mode* will include all three pins—H, L, and W—except in some instances where the low end of the ladder is grounded internally.

Variants are shown in Figures 8-7, 8-8, 8-9, and 8-10. Because some pins may be omitted, and there is no standardization of function among the pins that do exist, circuits and chips must be examined carefully prior to use.

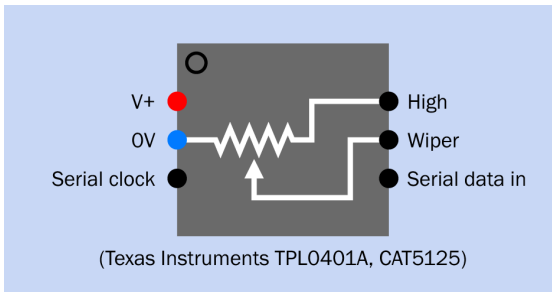


Figure 8-7. Some digital potentiometers minimize chip size and provide specialized functionality by eliminating pins. In the variant shown here, the W pin provides a voltage between H and an internal ground connection. The chip is controlled via I2C serial protocol.

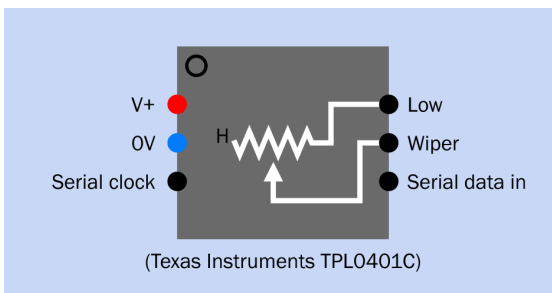


Figure 8-8. In this variant, the H end of the internal resistor ladder is allowed to float inside the chip, and the digital potentiometer functions as a rheostat. The chip is controlled via I2C serial protocol.

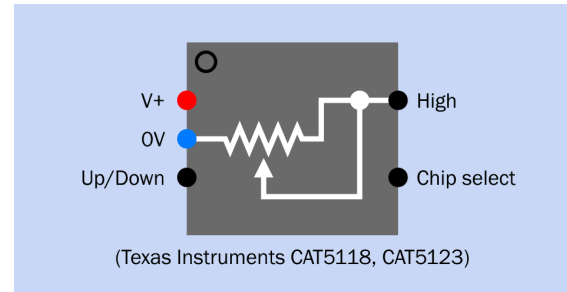


Figure 8-9. This variant provides a variable resistance between the H pin and an internal connection with negative ground. Pin 5 is omitted. The chip is controlled by up/down pulses.

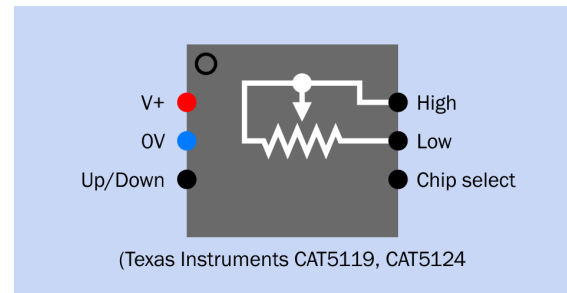


Figure 8-10. This variant provides a variable resistance between H and L pins, without allowing either end of the resistor ladder to float. The W pin is omitted, as the wiper is tied internally to the H pin. The chips listed are controlled by up/down pulses.

Values

A primary limitation of digital potentiometers is that they cannot withstand significant current. This may prevent them from being substituted for an analog potentiometer unless changes are made in the circuit. H, L, and W pins are usually unable to source or sink continuous, sustained current exceeding 20mA.

Wiper resistance is the resistance that is added internally by the wiper. This is nontrivial; it is often around 100Ω, and can be as high as 200Ω.

Typical end-to-end resistance of the ladder of internal resistors may range from 1K to 100K. Values of 1K, 10K, and 100K are common.

While the number of taps is likely to be a power of two in chips where the taps are addressable, up/down chips are not so constrained and may contain, for example, 100 taps.

The end-to-end resistance of a whole ladder may vary by as much as 20% from one sample of a chip to the next. Among resistor ladders in digital potentiometers sharing the same chip (i.e., in dual or quad chips) the variation will be much smaller.

Almost all digital potentiometers are designed for a supply voltage of 5V or less. The H and L pins are not sensitive to polarity, but the voltage applied to either of them must not exceed the supply voltage.

How to Use It

While most microcontrollers contain one or more analog-digital converters that change an analog *input* to an internal numeric value, a microcontroller cannot create an analog *output*. A digital potentiometer adds this functionality, although applications will be restricted by its limitation on current.

An up/down digital potentiometer can be controlled directly by a pair of **pushbuttons**, one of which will increase the resistance value while the other will reduce it. The pushbuttons must be *debounced* when used in this way. An alternative to pushbuttons is a **rotational encoder**, which emits a stream of pulses when its shaft is turned. In this case, an intermediate component (probably a microcontroller) will usually interpret the pulse stream and change it to a format that the digital potentiometer can understand.

Where a digital potentiometer is used in audio applications, it should be of the type that moves the wiper connection from one tap to another during a *zero crossing* of the audio signal (i.e., that is, at the moment when the AC input signal passes through 0V on its way from positive to negative or negative to positive). This suppresses the “click” that otherwise tends to occur during

switching. Potentiometers with this feature may include phrases such as “glitch free” in their datasheets.

Digital potentiometers that are intended primarily for audio applications often have 32 taps spaced at intervals of 2dB. This will be sufficient to satisfy most listeners.

Achieving Higher Resolution

For sensitive applications where a resolution with more than 1,024 steps is required, multiple digital potentiometers with different step values can be combined. One way of doing this is shown in [Figure 8-11](#). In this circuit, the wipers of P2 and P3 must be moved in identical steps, so that the total resistance between the positive power supply and negative ground remains constant. These two potentiometers could be contained in a dual chip, and would receive identical up/down commands. P1 is at the center of the voltage divider formed by P2 and P3, and is adjusted separately to “fine tune” the output voltage that is sensed at point A.

If all three of the potentiometers in this circuit contain 100 taps, a combined total of 10,000 resistance steps will be possible.

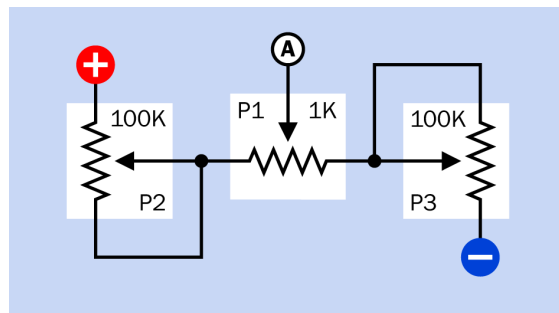


Figure 8-11. If all three digital potentiometers in this schematic have 100 taps, and the wipers of P2 and P3 are moved in synchronization, the voltage measured at point A can have a high resolution of up to 10,000 steps.

What Can Go Wrong

Noise and Bad Inputs

Because a digital potentiometer is capable of receiving data at speeds as high as 1MHz, it is sensitive to brief input or power fluctuations, and can misinterpret them as instructions to move the wiper—or can misinterpret them as command codes, in a component using SPI or I2C serial protocol.

To minimize noise in the power supply, some manufacturers recommend installing a 0.1µF capacitor as close as possible to the power supply pin of the component. In addition, it is obviously important to provide clean input signals. This means thorough debouncing of any electromechanical switch or pushbutton inputs.

Wrong Chip

The wide diversity of input protocols and pinouts creates many opportunities for installation error.

Up/down, SPI, and I2C protocol require totally different pulse streams. Many manufacturers offer components that are distinguished from each other by just one or two digits in their part numbers, yet have radically different functionality.

If more than one specific type of digital potentiometer may be used during circuit development, they should be stored carefully to avoid inadvertent substitutions. Using the wrong chip may be particularly confusing in that an inappropriate input protocol will still produce some results, although not those which were intended.

Controller and Chip Out of Sync

As noted in the discussion of data transmission protocols, most digital potentiometers are not

capable of providing feedback to confirm the position of the internal wiper. A designer may wish to include a power-up routine which establishes the state of the digital potentiometer by resetting it to a known position, at one end of its scale or the other.

Nonlinear Effects

While the end-to-end resistance of the resistor ladder inside a digital potentiometer is not likely to be affected significantly by changes in temperature, the resistance at the wiper is more heat sensitive.

In an up/down chip, there can be differential errors between incremental and decremental modes. In other words, if a tap is reached by stepping up to it incrementally, the resistance between the W pin and H or L may not be quite the same as if the same tap is reached by stepping down to it decrementally. The difference may not be significant, but may be puzzling for those who are unfamiliar with this phenomenon.

Some differences may be found among resistors in a ladder. That is, in a supposedly linear digital potentiometer, each resistor may differ in value slightly from the next.

Data Transfer Too Fast

When using a microcontroller to send data to a digital potentiometer, a small delay may be necessary between pulses, depending on the microcontroller's clock speed. A digital potentiometer may require a minimum pulse duration of 500ns. Check the manufacturer's datasheet for details.

timer



A device that creates a single timed pulse, or a series of timed pulses with timed intervals between them, is properly known as a *multivibrator*, although the generic term **timer** has become much more common and is used here.

Three types of multivibrator exist: *astable*, *monostable*, and *bistable*. The behavior of *astable multivibrators* and *monostable multivibrators* is described in detail in this entry. A timer chip can also be made to function as a bistable multivibrator. This is described briefly below, but it is not a designed function of a timer. The primary discussion of bistable multivibrators will be found in the entry of this encyclopedia dealing with **flip-flops**.

OTHER RELATED COMPONENTS

- **flip-flop** (See [Chapter 11](#))

What It Does

A *monostable* timer emits a single timed pulse of fixed length in response to a triggering input that is usually of shorter duration. Many monostable timers are also capable of running in *astable* mode, in which the timer spontaneously emits an ongoing stream of timed pulses with timed gaps between them. A dual-mode timer can run in either mode, determined either by external components attached to it, or (less commonly) by changing the status of a mode selection pin.

Monostable Mode

In monostable mode, the timer emits a pulse in response to a change from high to low voltage (or, less commonly, from low to high voltage) at a *trigger pin*. Most timers respond to a *voltage level* at the trigger pin, but some are insensitive to any persistent pin state and only respond to a *voltage transition*. This is known as *edge triggering*.

The pulse generated by the timer may consist of a change from low to high (or, less commonly, from high to low) at an output pin. The length of the pulse will be determined by external components, and is independent of the duration of the triggering event, although in some cases, an output pulse may be prolonged by *retriggering* the timer prematurely. This is discussed below.

At the end of the output pulse, the timer reverts to its quiescent state, and remains inactive until it is triggered again.

A monostable timer can control the duration of an event, such as the time for which a light remains on after it has been triggered by a motion sensor. Alternatively, the timer can impose a delay, such as the interval during which a paper towel dispenser refuses to respond after a towel has been dispensed. A timer can also be useful to generate a clean pulse in response to an unstable or noisy input, as from a manually operated pushbutton.

Astable Mode

In astable mode, a timer will generally trigger itself as soon as power is connected, without any need for an external stimulus. However, the output can be suppressed by applying an appropriate voltage to a *reset pin*.

External components will determine the duration of each pulse and the gap between it and the next pulse. The pulse stream can be slow enough to control the flashing of a turn signal in a 1980s automobile, or fast enough to control the bit rate in a data stream from a computer.

Modern timer circuits are often incorporated in chips that have other purposes. The flashing of a turn signal in a modern car, for instance, is now likely to be timed by a microcontroller that manages many other functions. Still, chips that are purely designed as timers remain widely used and are very commonly available in numerous through-hole and surface-mount formats.

How It Works

The duration of a single pulse in monostable mode, or the frequency of pulses in astable mode, is most commonly determined by an external *RC network* consisting of a resistor in series with a capacitor. The charging time of the capacitor is determined by its own size and by the value of the resistor. The discharge time will be determined in the same way. A **comparator** inside the timer is often used to detect when the potential on the capacitor reaches a *reference voltage* that is established by a voltage divider inside the chip.

Variants

The 555 Timer

An eight-pin integrated circuit manufactured by Signetics under part number 555 was the world's first fully functioned timer chip, introduced in 1972. It combined two comparators with a **flip-flop** (see [Chapter 11](#)) to enable great versatility while maintaining excellent stability over a wide range of supply voltages and operating temper-

atures. Subsequent timers have been heavily influenced by this design. A typical 555 timer chip is shown in [Figure 9-1](#).

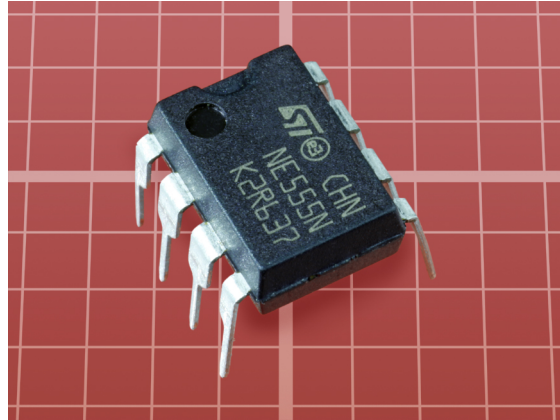


Figure 9-1. A typical 555 timer chip. Functionally identical versions in which the “555” identifier is preceded or followed by different letter combinations are available from many different manufacturers.

The 555 was designed by one individual, Hans Camenzind, working as an independent consultant for Signetics. According to a transcript of an interview with Camenzind maintained online at the Transistor Museum, “There was nothing like it at the time. You had to use quite a few discrete components—a comparator, a Zener diode or even two. It was not a simple circuit.”

The 555 timer quickly became the most widely used chip in the world, and was still selling an annual estimated 1 billion units three decades after its introduction. It has been used in spacecraft, in intermittent windshield wiper controllers, in the early Apple II (to flash the cursor), and in children’s toys. Like many chips of its era, its design was unprotected by patents, allowing it to be copied by numerous manufacturers.

The initial version was built around **bipolar transistors**, and consequently is referred to as the *bipolar version* or (more often) the *TTL version*, this being a reference to *transistor-transistor logic* protocol. Within a few years, CMOS versions based around **MOSFETs** were developed. They reduced the ability of the chip to sink or source

current at its output pin, but consumed far less power, making them better suited to battery-operated products. The CMOS versions were and still are pin-compatible with the original bipolar version, both in through-hole and surface-mount formats. Their timing parameters are usually the same.

555 Monostable Operation

The internal functionality of a 555 timer wired to run in monostable mode is illustrated in Figure 9-2 with the chip seen from above. The pins are identified in datasheets by the names shown. To assist in visualizing the behavior of the chip, this figure represents the internal flip-flop as a switch which can be moved by either of two internal comparators, or by an input from the Reset pin.

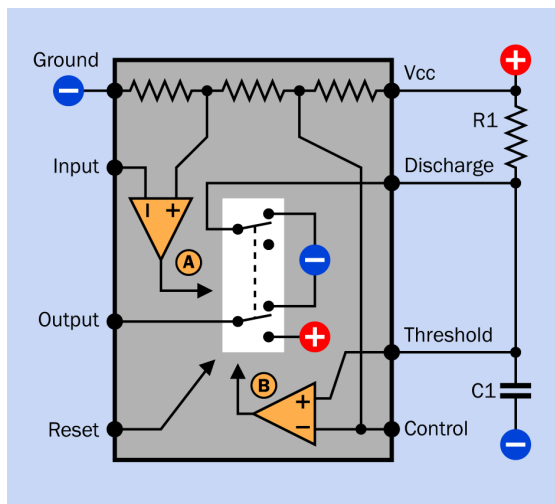


Figure 9-2. The internal functions of a 555 chip, with its flip-flop represented as a switch that can be moved by either of two comparators, or by a low voltage on the Reset pin. An external resistor and capacitor, shown as R1 and C1, cause the timer to run in monostable (one-shot) mode, generating a single high pulse when the state of the Input pin is pulled from high to low.

Inside the chip, three resistances of 5K each are connected between V+ (positive supply voltage) and negative ground. It has been suggested that the part number of the 555 chip was derived from these three 5K resistors, but Hans Camenzind has

pointed out that Signetics was already using three-digit part numbers beginning with the number 5, and probably chose the 555 part number because the sales department had high expectations for the chip and wanted its number to be easily memorable. (A similar rationale explains the part number of the 2N2222 transistor.)

The resistances inside the timer function as a *voltage divider*, providing a reference of $1/3$ of V+ to the noninverting pin of Comparator A and $2/3$ of V+ to the inverting pin of Comparator B. (See Chapter 6 for an explanation of the functioning of comparators.)

When power is initially supplied to the timer, if the Input pin is at a high state, Comparator A has a low output, and the flip-flop remains in its “up” position, allowing the Output pin to remain in a low state. The flip-flop also grounds the lower end of R1, which prevents any charge from accumulating on capacitor C1.

If the state of the Input pin is pulled down externally to a voltage less than $1/3$ of V+, Comparator A now creates a high output that changes the flip-flop to its “down” position, sending a high signal out through the Output pin. At the same time, C1 is no longer grounded, and begins to charge at a rate determined by its own size and by the value of R1. When the charge on the capacitor exceeds $2/3$ of V+, it activates Comparator B, which forces the flip-flop into its “up” position. The Output pin goes low, C1 discharges itself into the Discharge pin, and the timer’s cycle is at an end.

The low voltage on the Input pin of the timer must end before the end of the output cycle. If the voltage on the Input pin remains low, it will re-trigger the timer, prolonging the output pulse.

A pullup resistor may be used on the Input pin to avoid false triggering, especially if an external electromechanical **switch** or **pushbutton** is used to pull down the Input pin voltage.

The Reset pin should normally be held high, either by being connected directly to positive supply voltage (if the reset function will not be needed) or by using a pullup resistor. If the Reset pin is pulled low, this will always interrupt an output pulse regardless of the timer's current status.

If a voltage higher or lower than $2/3$ of $V+$ is applied to the Control pin, this will change the reference voltage on Comparator B, which determines when the charging cycle of C1 ends and the discharge cycle begins. A lower reference voltage will shorten each output pulse by allowing a lower charge limit for C1. If the control voltage drops to $1/3$ of $V+$ (or less), the capacitor will not charge at all, and the pulse length will diminish to zero. If the control voltage rises to become equal to $V+$, the capacitor will never quite reach that level, and the pulse length will become infinite. A workable range for the control voltage is therefore 40% to 90% of $V+$.

Because the Control pin is an input to the chip, it should be grounded through a $0.01\mu\text{F}$ ceramic capacitor if it will not be used. This is especially important in CMOS versions of the timer.

A defect of the bipolar 555 is that it creates a *voltage spike* when its Output pin changes state. If it will be sharing a circuit with sensitive components, a $0.01\mu\text{F}$ bypass capacitor should be added as closely as possible between the $V+$ pin and negative ground. The voltage-spike problem was largely resolved by the CMOS 555.

555 Astable Operation

In [Figure 9-3](#), the 555 timer chip is shown with external components and connections to run it in astable mode. The pin names remain the same but have been omitted from this diagram because of limited space. The labeling of the two external resistors and capacitor as R1, R2, and C1 is universal in datasheets and manufacturers' documentation.

When the timer is powered up initially, capacitor C1 has not yet accumulated any charge. Consequently, the state of the Threshold pin is low. But

the Threshold pin is connected externally with the Input pin, for astable operation. Consequently, the Input pin is low, which forces the flip-flop into its "down" state, creating a high output. This happens almost instantaneously.

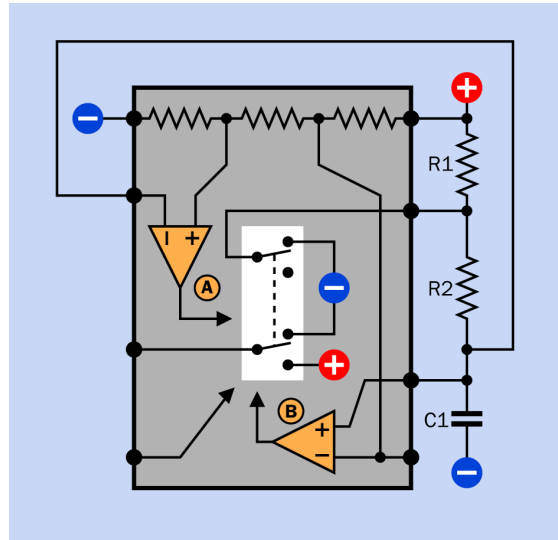


Figure 9-3. The internal functions of a 555 chip, with two external resistors and a capacitor wired to run the timer in astable (free-running) mode.

While the flip-flop is "down," the Discharge pin is not grounded, and current flowing through R1 and R2 begins to charge the capacitor. When the charge exceeds $2/3$ of positive supply voltage, Comparator B forces the flip-flop into its "up" position. This ends the high pulse on the Output pin, and starts to drain the charge from the capacitor through R2, into the Discharge pin. However, the voltage on the capacitor is still being shared by the Input pin, and when it diminishes to $1/3$ of $V+$, the Input pin reactivates Comparator A, starting the cycle over again.

The functions of the Reset and Control pins are the same as in monostable mode. Because voltage applied to the Control pin changes the length of each pulse and the gaps between pulses, it has the effect of adjusting the frequency of the output in astable mode.

When power is first connected to the timer, C1 must initially charge from an assumed state of zero potential to $2/3 V+$. Because subsequent cycles will begin when the capacitor is at $1/3 V+$, the first high output pulse from the timer will be slightly longer than subsequent output pulses. This is unimportant in most applications, especially because the rate at which a capacitor accumulates charge is greater when beginning from 0V than when it has reached $1/3 V+$. Still, the longer initial pulse can be noticeable when the timer is running slowly.

Because the capacitor charges through R1 and R2 in series, but discharges only through R2, the length of each positive output pulse in astable mode is always greater than the gap between pulses. Two strategies have been used to overcome this limitation. See [“Separate Control of High and Low Output Times”](#) on page 80.

556 Timer

The 556 consists of two 555 bipolar-type timers in one package. An example of the chip is shown in [Figure 9-4](#). The pinouts are shown in [Figure 9-5](#). Although 556 timers have become relatively uncommon compared with the 555, they are still being manufactured in through-hole and surface-mount versions by companies such as Texas Instruments and STMicroelectronics, under part numbers such as NA556, NE556, SA556, and SE556 (with various letters or letter pairs appended). Each timer in the chip has its own set of inputs and outputs, but the timers share the same $V+$ and ground voltages.

558 Timer

This 16-pin chip is now uncommon, and many versions have become obsolete. It has been identified by a part number such as NE558 although different prefix letters may be used. The NTE926, shown in [Figure 9-6](#), is actually a 558 timer.

The chip contains four 555 timers sharing a common power supply, common ground, and common control-pin input. For each internal timer, the Threshold and Discharge functions are con-

nected internally, so that the timers can only be used in one-shot mode. However, one timer can trigger another at the end of its cycle, and the second timer can then retrigger the first, to create an astable effect.

Each timer is edge-triggered by a voltage transition (from high to low), instead of being sensitive to a voltage level, as is the case with a 555 timer. Consequently the timers in the 558 chip are insensitive to a constant (DC) voltage.

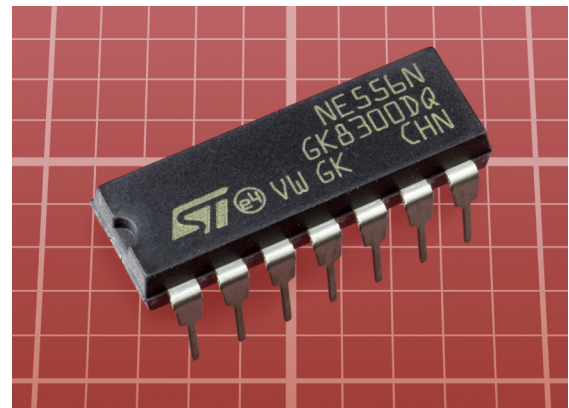


Figure 9-4. An example of the 556 timer chip.

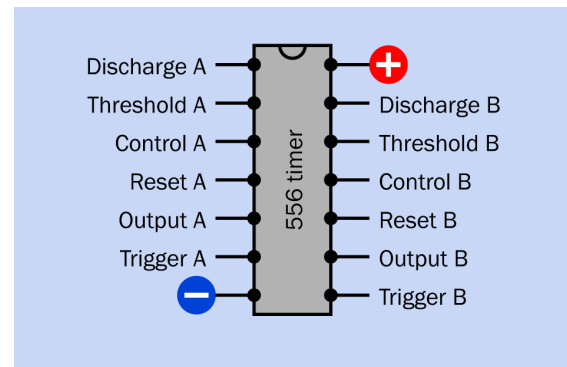


Figure 9-5. The 556 timer contains two separate 555 timers sharing the same power supply and ground. The pin functions for timer A and timer B are shown here.

The output from each timer is an open collector, and therefore requires an external pullup resistor. Each output is capable of sinking up to 100mA.

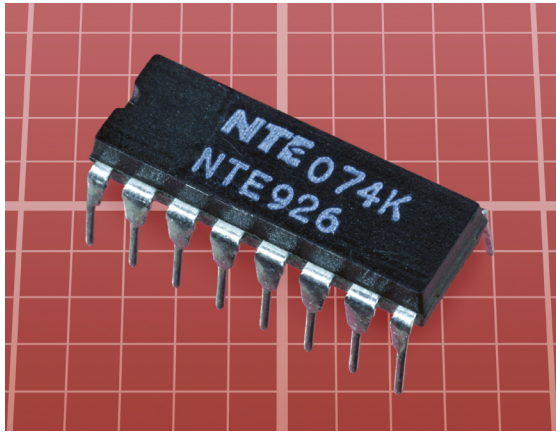


Figure 9-6. The NTE926 is a 558 timer chip.

CMOS 555 Timer

While the part numbers of many CMOS versions are significantly different from part numbers of the bipolar versions, in some instances the CMOS numbers are only distinguished by a couple of initial letters. The ST Microelectronics TS555 series and Texas Instruments TLC555 series, for example, use MOSFETs internally. The ST Microelectronics SE555 series and Texas Instruments SA555 series use bipolar transistors internally.

One way to distinguish between the two types, when searching a website maintained by a parts supplier, is to begin by looking generically for a “555 timer” and then add a search filter to show chips either with a minimum power supply of 3VDC (which will be CMOS) or with a minimum power supply of 4.5VDC (which will be bipolar).

CMOS versions of the 555 timer do not create the power spike that is characteristic of the bipolar versions during output transitions. The CMOS chips can also be powered by a lower voltage (3VDC, or 2VDC in some cases), and will draw significantly less current in their quiescent state. They also require very little current for threshold, trigger, and reset functions.

The wiring of external resistors and capacitors to the CMOS version of the chip, and the internal voltage levels as a fraction of V_{+} , are identical to

the original 555 timer. Pin functions are likewise identical. The only disadvantages of CMOS versions are their greater vulnerability to static discharge, and their lower output currents. The TLC555, for instance, will source only 15mA (although it can sink 10 times that amount). Other manufacturers have different specifications, and datasheets should be checked carefully.

5555 Timer

The 5555 contains a digital **counter** that enables it to time very long periods. Its full part number is 74HC5555 or 74HCT5555, although these numbers may be preceded or followed by letter combinations identifying the manufacturer. It is not pin-compatible with a 555 timer.

Two input pins are provided, one to trigger the timer on a rising edge, the other on a falling edge, of the input pulse. The inputs are Schmitt-triggered.

The timer is rated for 1Hz to 1MHz (using an external resistor and external capacitor). The counter section can divide the pulse frequency by values ranging from 2 to 256. For longer timed periods, different settings on the digital control pins will divide the frequency by values ranging from 2^{17} through 2^{24} (131,072 through 16,777,216). This enables the timer to achieve a theoretical pulse length lasting for more than 190 days. The timer will accept a clock input from an external oscillator to achieve better accuracy than is available with a resistor-capacitor timing circuit.

7555 Timer

This 8-pin chip is a CMOS version of the 555 timer, manufactured by companies such as Maxim Integrated Products and Advanced Linear Devices. Its characteristics are similar to those of CMOS 555 timers listed above, and the pinouts are the same.

7556 Timer

This 14-pin chip contains two 7555 timers, sharing common power supply and ground connec-

tions. Pinouts are the same as for the original 555 timer, as shown in [Figure 9-5](#).

4047B Timer

This 14-pin CMOS chip was introduced in an effort to address some of the quirks of the 555 timer while also providing additional features. It runs in either monostable or astable mode, selectable by holding one input pin high or another input pin low. In astable mode, its duty cycle is fixed at approximately 50%, a single resistor being used for both charging and discharging the timing capacitor. An additional “oscillator” output runs twice as fast as the regular output.

In monostable mode, the 4047B can be triggered by a positive or negative transition (depending on which of two input pins is used). It ignores steady input states and will also ignore additional trigger pulses that occur during the output pulse. However, a retrigger pin is provided to extend the output pulse if desired.

Complementary output pins are provided, one being active-high while the other is active-low.

To time very long periods, the 4047B was designed to facilitate connection with an external counter.

The power supply for the 4047B can be as low as 3VDC. Its maximum source or sink output current is only 1mA when powered at 5VDC, but up to 6.8mA when powered at 15VDC.

The chip is still available from manufacturers such as Texas Instruments (which markets it as the CD4047B) in through-hole and surface-mount formats. However, despite its versatility, the 4047B is less popular than dual monostable timers, described in the next section.

Dual Monostable Timers

Various timers that run only in monostable mode are available in dual format (i.e., two timers in one chip). This format became popular partly because two monostable timers can trigger each other to create an astable output, in which the pulse width, and the gap between pulses, can be

set by a separate resistor-capacitor pair on each timer. This allows greater flexibility than is available when using a 555 timer.

Most dual monostable timer chips are edge-triggered by a change in input voltage, and will ignore a steady DC voltage. Consequently, the output from one timer can be connected directly to the input of another, and no coupling capacitors are necessary.

As in the 4047B, the user has a choice of two input pins for each timer, one triggered by a transition from low to high, the other triggered by a transition from high to low. Similarly, each timer has two outputs, one shifting from low to high at the start of the output cycle, the other shifting from high to low.

The values of a single resistor and capacitor determine the pulse duration of each timer.

Dual monostable timers often have the numeric sequence 4528 or 4538 in their part numbers. Examples include the HEF4528B from NXP, the M74HC4538 from STMicroelectronics, and the MC14538B from On Semiconductor. The 74123 numeric sequence identifies chips that have a very similar specification, with chip-family identifiers such as HC or LS inserted, as in the 74HC123 and 74LS123, and additional letters added as prefix or suffix. The pinouts of almost all these chips are identical, as shown in [Figure 9-7](#). However, Texas Instruments uses its own numbering system, and datasheets should always be consulted for verification before any connections are made.

Many chips of this type are described as “retriggerable,” meaning that if an additional trigger pulse is applied to the input before an output pulse has ended, the current output pulse will be extended in duration. Check datasheets carefully to determine whether a chip is “retriggerable” or will ignore new inputs during the output pulse.

The 74HC221 dual monostable vibrator (pictured in [Figure 9-8](#)) functions very similarly to the

components cited above, but has slightly different pinouts.

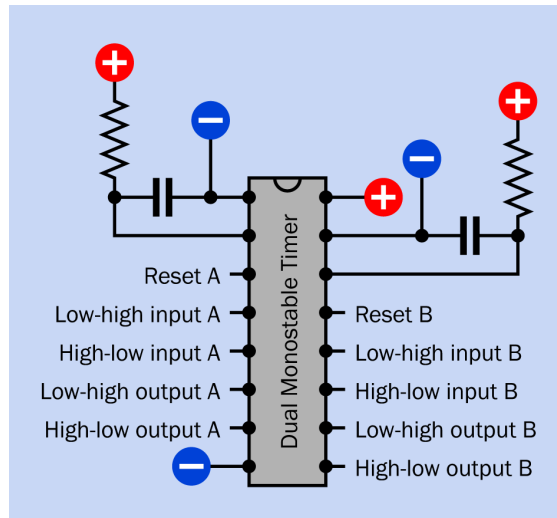


Figure 9-7. Pin functions for most 4528, 4538, and 74123 series of dual monostable timer chips. An RC network is shown connected for each timer. Note that Texas Instruments uses different pinouts on its versions.

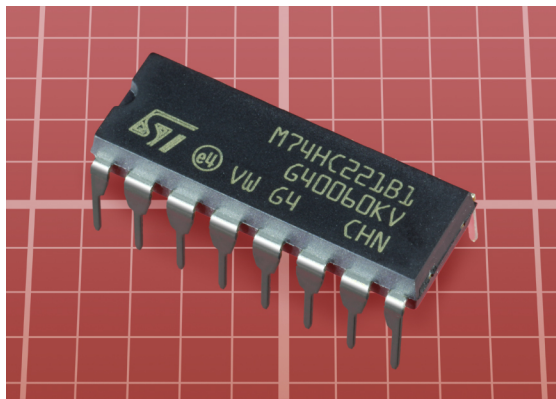


Figure 9-8. A dual-timer chip containing two monostable multivibrators that can function in astable mode if they are connected externally to trigger each other.

Values

555 Timer Values

The original bipolar version of the 555 timer was designed to operate using a wide range of positive supply voltages, from 4.5VDC to 16VDC.

CMOS versions vary in their recommended $V+$ values, and datasheets must be consulted for verification.

The output of a bipolar 555 is rated to source or sink up to 200mA. In practice, the maximum current will be lower when the timer is powered at the low end of its range, around 5VDC. Attempting to source more than 50mA will pull down the voltage internally, affecting operation of the timer.

CMOS versions all impose restrictions on output current, allowing higher values for sinking than sourcing. Again, datasheets must be consulted for the values, which vary widely from one component to another.

The voltage measured on the output pin, when it is used for sourcing current, will always be lower than the power supply voltage, and a 1.7V drop is commonly specified for bipolar versions. In practice, the voltage drop that is actually measured may be less, and will vary according to the load on the output.

The voltage drop does not increase significantly with a higher supply voltage, and because it is a relatively constant value, it becomes less significant when a higher value for $V+$ is used.

CMOS versions of the 555 timer achieve a claimed output source voltage that is only 0.2V less than the power supply.

When choosing values for $R1$ and $R2$, a minimum for each resistor is 5K, although 10K is preferred. Lower values will increase power consumption, and may also allow overload of the internal electronics when the chip sinks current from $C1$. A typical maximum value for each resistor is 10M.

A high-value capacitor may cause the timer to function less accurately and predictably, because large capacitors generally allow more *leakage*. This means that the capacitor will be losing charge at the same time that it is being charged through $R1 + R2$. If these resistors have high values, and the capacitor has a value of 100 μ F or more, the rate of charge may be so low that it is

comparable with the rate of leakage. For this reason, a 555 timer is not a good choice for timing intervals much greater than a minute. If a large-value capacitor is used, tantalum is preferable to electrolytic.

The minimum practical value for a timing capacitor is around 100pF. Below this, performance may not be reliable.

Although some CMOS versions may enable fast switching, the shortest practical output pulse for a 555 timer is around 10 microseconds. On the input pin, a triggering pulse of at least 1 microsecond should be used.

Time Calculation in Monostable Mode

If R1 is measured in kilohms and C1 is measured in microfarads, the pulse duration, T, in seconds, of a 555 timer running in monostable mode can be found from this simple formula:

$$T = 0.0011 * R1 * C1$$

This relationship is the same in all versions of the 555. [Figure 9-9](#) provides a quick and convenient way to find the pulse value using some common values for R1 and C1. Resistors can be obtained with tolerances below plus-or-minus 1%, but capacitors are often rated with an accuracy of only plus-or-minus 20%. This will limit the accuracy of the pulse values shown in the chart.

Time Calculation in Astable Mode

If R1 and R2 are measured in kilohms and C1 is measured in microfarads, the frequency of pulses, F (measured in Hz) of a 555 timer running in astable mode can be found from this simple formula:

$$F = 1440 / ((R1 + (2 * R2)) * C1)$$

This relationship is found in all versions of the 555. [Figure 9-10](#) shows the frequency for common values of R2 and C1, assuming that the value of R1 is 10K. In [Figure 9-11](#), a value of 100K is assumed for R1.

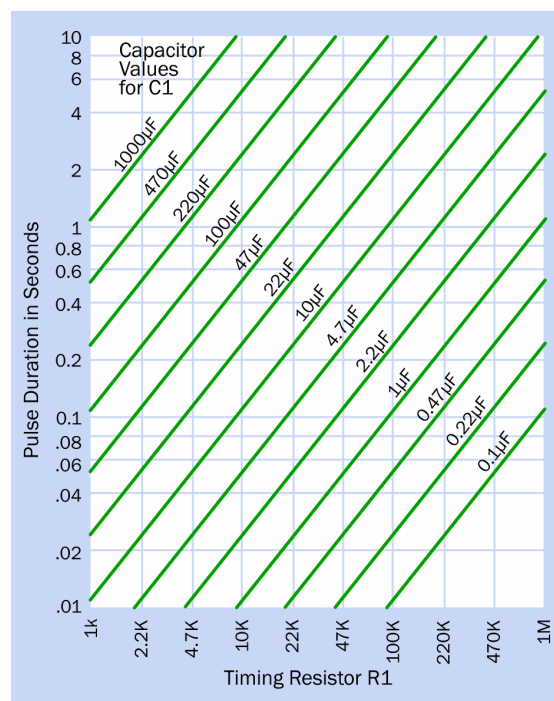


Figure 9-9. To determine the pulse duration of a 555 timer running in monostable mode: find the value of R1 on the horizontal scale, follow its vertical grid line upward to the intersection with a green line which corresponds with the value of capacitor C1, and read across to the vertical scale providing the duration in seconds. Both axes are logarithmic.

Dual Monostable Timers

Dual chips such as the HEF4528B from NXP, the M74HC4538 from STMicroelectronics, the MC14538B from On Semiconductor, and the 74HC123 from Texas Instruments have widely varying requirements for power supply. Some accept a limited range from 3VDC to 6VDC, while others tolerate a range of 3VDC to 20VDC. When powered with 5VDC their required input and output states are compatible with those of 5V logic chips.

Output pins of these chips source and sink no more than 25mA (much less in some instances). Because there are so many variants, they cannot be summarized here, and datasheets must be consulted for details.

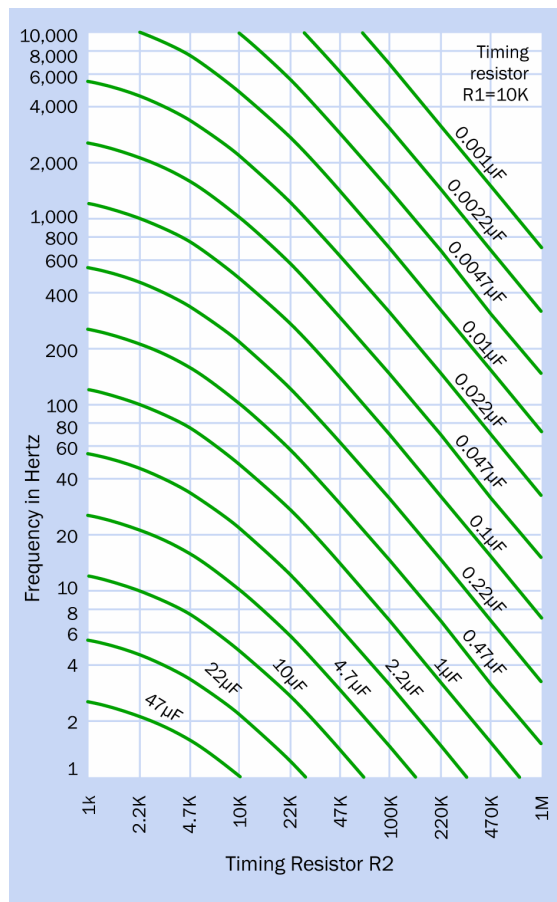


Figure 9-10. To obtain the frequency of a 555 timer running in astable mode, when R1 has a value of 10K: find the value of R2 on the horizontal scale, follow its vertical grid line upward to the intersection with a green curve which corresponds with the value of capacitor C1, and read across to the vertical scale providing the frequency in Hertz. Both axes are logarithmic.

As these timers are all monostable, and each timer uses just one resistor and one capacitor, the only formula required is to give the pulse time as a function of these two variables. If R is the resistor value in ohms, and F is the capacitor value in farads, and K is a constant supplied by the manufacturer, the pulse time T, in seconds, is found from the formula:

$$T = R * F * K$$

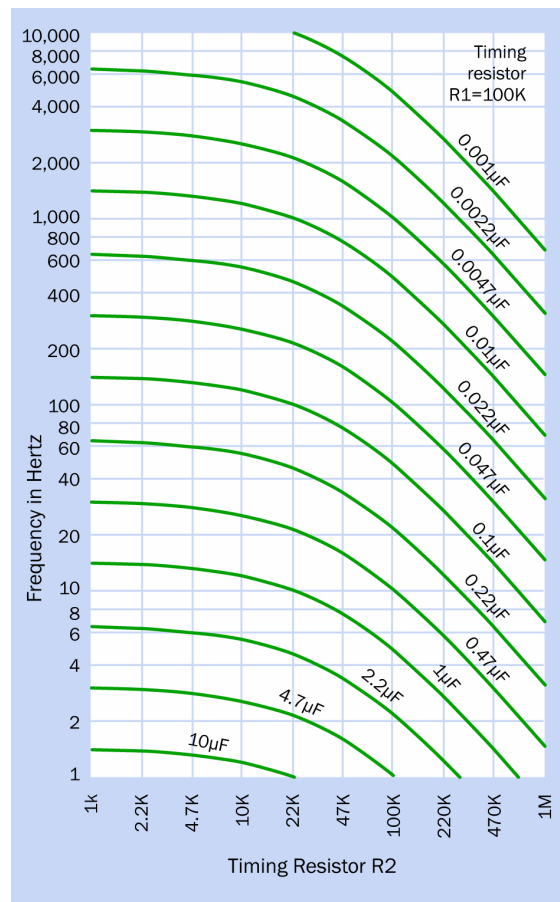


Figure 9-11. To obtain the frequency of a 555 timer running in astable mode, when R1 has a value of 100K: find the value of R2 on the horizontal scale, follow its vertical grid line upward to the intersection with a green curve which corresponds with the value of capacitor C1, and read across to the vertical scale providing the frequency in Hertz. Both axes are logarithmic.

K ranges between 0.3 and 0.7 depending on the manufacturer and also on the voltage being used. Its value should be found in the manufacturer's datasheet. If R is measured in megohms and F is measured in microfarads, the formula is still valid, as the multipliers cancel each other out.

Generally speaking, these dual monostable CMOS timers are not intended for pulse duration exceeding 1 minute.

The timing capacitor should be no larger than $10\mu\text{F}$, as it discharges directly and rapidly through the chip.

How to Use It

Where a timer is required to drive a load such as a relay coil or small motor directly, the original TTL version of the 555 timer will be the only choice. Even in this instance, a protection diode must be used across the inductive device.

For smaller loads and applications in chip-to-chip circuits, CMOS versions of the 555, including the 7555, use less power, cause less electrical interference, and are pin-compatible while using the same formulae to calculate frequency in astable mode or pulse duration in monostable mode. They are of course more vulnerable to static discharge, and care must be taken to make a connection to every pin (the capacitor that grounds the Control pin, if Control is not going to be used, is mandatory).

In dual monostable timers, unused rising-edge trigger inputs should be tied to V+ while unused falling-edge trigger inputs should be tied to ground. A Reset pin that will not be used should be tied to V+, unless that entire timer section of the chip will be unused, in which case the pin should be grounded.

To measure durations longer than a few minutes, a timer which incorporates a programmable counter to divide the clock frequency is the sensible choice. See the description of the 5555 timer that was included earlier in this entry.

The original bipolar version of the 555 remains a robust choice in hobby applications such as robotics, and its design allows some versatile variations which may even be used in logic circuits. A variety of configurations are shown in the schematics below.

555 Monostable Mode

The basic schematic for a 555 timer running in monostable mode is shown in [Figure 9-12](#). In this

particular example, a pushbutton that is liable to suffer from *switch bounce* is connected to the Input pin of the timer, which responds to the very first connection made by the pushbutton and ignores the subsequent “bounces,” thus producing a “clean” output. To avoid retriggering, which results in a prolonged output pulse, the timer’s output should exceed the time for which the button is likely to be pressed. The output should also exceed the duration of any possible switch bounce, which can otherwise create multiple output pulses. In the schematic, an LED is attached to the timer output for demonstration purposes.

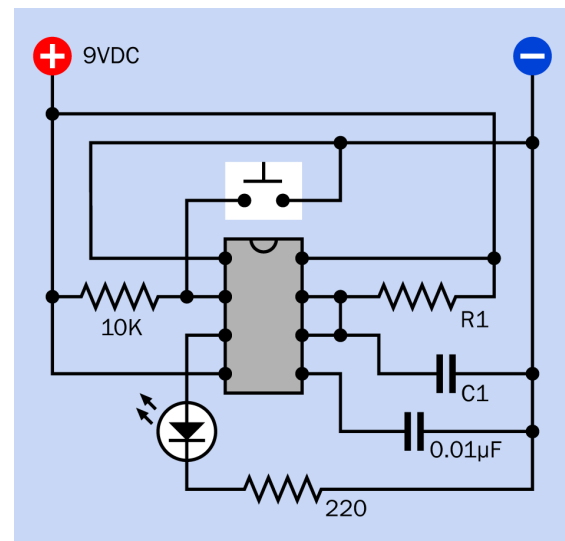


Figure 9-12. The basic monostable configuration of a 555 timer. This particular circuit debounces an input from a pushbutton switch and converts it to a clean pulse of fixed duration, powering an LED for demonstration purposes.

This circuit is shown on a breadboard in [Figure 9-13](#). The red and blue wires, at the top of the photograph, supply 9VDC to the board. R1 is 1M , while C1 is $1\mu\text{F}$, creating a pulse of just over 1 second. A tactile switch, just above the timer, provides the input.

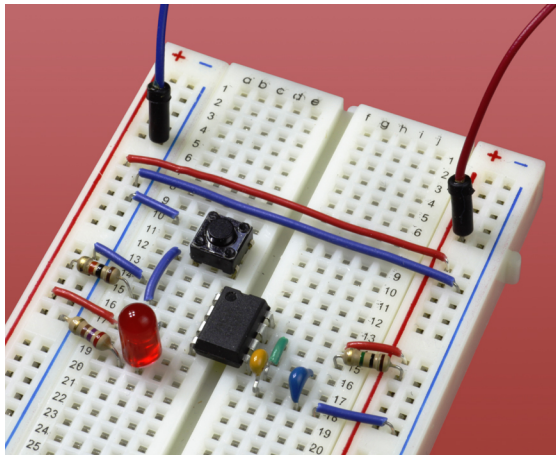


Figure 9-13. The basic monostable configuration for a 555 timer, mounted on a breadboard.

555 Astable Mode

A basic schematic for a 555 timer running in astable mode is shown in [Figure 9-14](#). Once again, an LED is attached to the output for demonstration purposes. If the pulse rate exceeds the persistence of vision, a small loudspeaker can be used instead, in series with a 47Ω resistor and a $100\mu\text{F}$ capacitor.

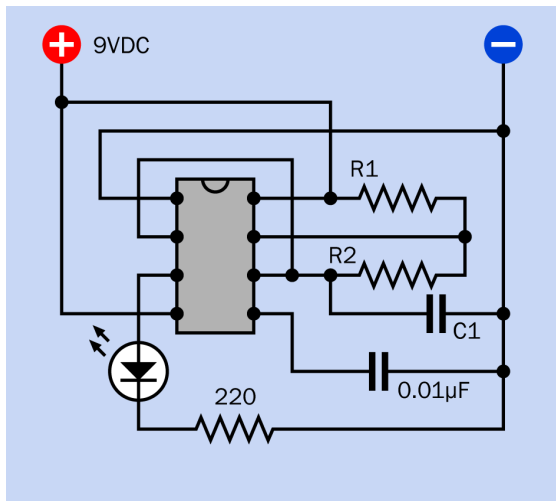


Figure 9-14. A 555 timer with external connections and components causing it to run in astable (free-running) mode.

Separate Control of High and Low Output Times

In [Figure 9-15](#), a bypass diode has been added around R2. The capacitor now charges primarily through R1, as the diode has a much lower effective resistance than R2. It discharges only through R2, as the diode blocks current in that direction. Consequently, the length of the high output pulse can be adjusted with the value of R1 only, while the length of the low output pulse can be adjusted with the value of R2 only. The duration of the high output can be lower than, or equal to, the duration of the low output, which is not possible with the basic configuration of components in [Figure 9-14](#).

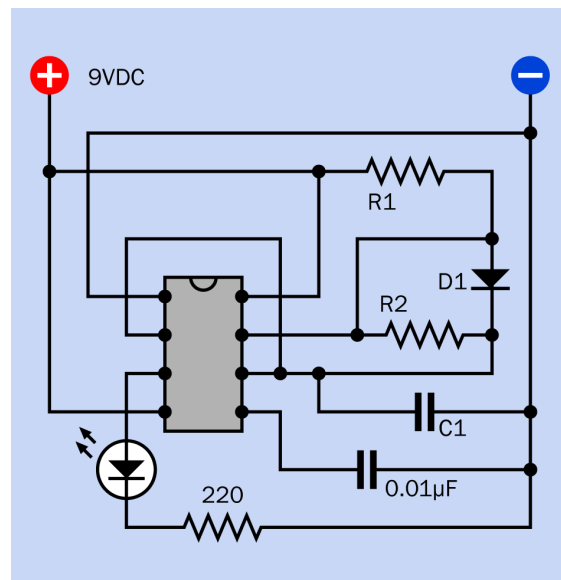


Figure 9-15. In this circuit, a diode bypasses R2, so that the “on” time and the “off” time of the 555 timer can be set independently of each other, with R1 and R2, respectively.

555 Fifty Percent Astable Duty Cycle: 1

In [Figure 9-16](#), the circuit enables a fixed astable output duration of approximately 50% high and 50% low. Initially, C1 has no charge, pulling the Input of the timer low, and causing it to begin a cycle with a high pulse from the Output pin, as

usual. In this demonstration circuit, the output illuminates an LED. At the same time, resistor R1 is attached to the output and charges C1. When the voltage on C1 reaches $2/3$ of $V+$, this is communicated to the timer Input pin, which ends the “high” cycle and initiates low status on the Output pin. This starts to sink the charge from C1, through R1. When the voltage drops to $1/3$ $V+$, this initiates a new cycle. Because only one resistor is used to charge and discharge the capacitor, we may imagine that the charge and discharge times should be identical. However, a higher load on the output will probably pull down the output voltage to some extent, lengthening the charge time. Conversely, a load on the Output pin that has low resistance will probably sink at least some of the charge from the capacitor, shortening the discharge cycle.

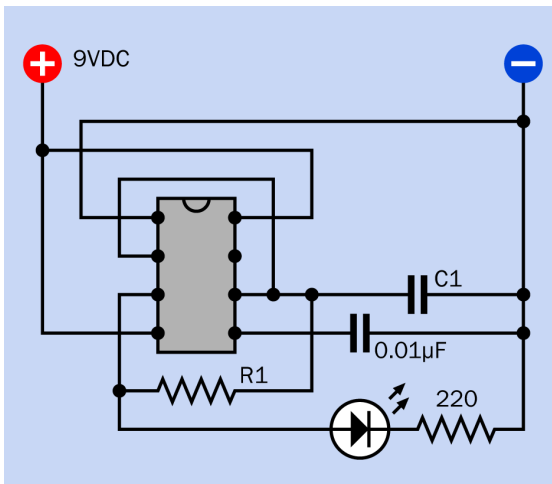


Figure 9-16. This configuration provides an approximate 50-50 on-off duty cycle at the output pin, although the precise duration will depend on the load.

555 Fifty Percent Astable Duty Cycle: 2

In [Figure 9-17](#), a small modification of the basic astable circuit shown in [Figure 9-14](#) provides another way to enable a 50% duty cycle. Compare the two schematics, and you will see that just the connection between R1 and R2 has been altered so that C1 now charges only through R1, and dis-

charges only through R2. However, in this configuration the capacitor is discharging into a voltage divider created by the two resistors. Empirical adjustment of the resistor values may be necessary before the duty cycle is precisely 50%.

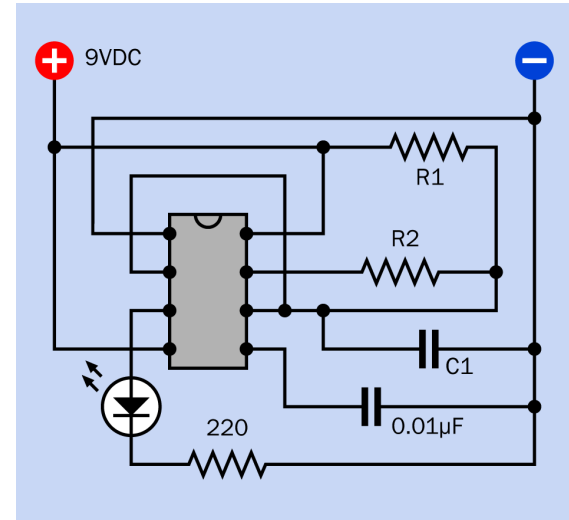


Figure 9-17. An alternative configuration to provide an approximate 50-50 on-off duty cycle in a 555 timer.

Use of the 555 Control Pin

In [Figure 9-18](#), a potentiometer and two series resistors allow a varying voltage to be applied to the Control Pin. This will lengthen or reduce both the charge and the discharge times of the timing capacitor. If values for the capacitor and its associated resistors are chosen to create a frequency of approximately 700Hz, a 10K potentiometer should demonstrate more than an octave of audible tones through the loudspeaker. Other components can be substituted for a potentiometer, creating possibilities for producing pulse-width modulation. Alternatively, if a large capacitor is added between the Control pin and ground while a second 555 timer, running slowly in astable mode, applies its output to the Control pin, the charging and discharging of the capacitor will apply a smoothly rising and falling voltage. If the first 555 timer is running at an audio frequency, the output will have a “wailing siren” effect.

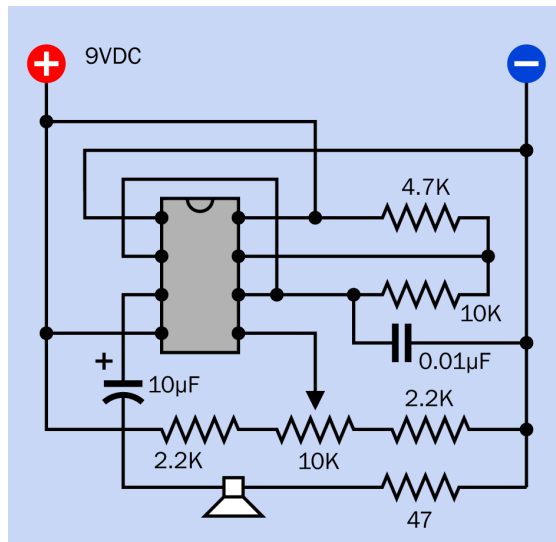


Figure 9-18. A circuit that allows adjustment of the as-table 555 frequency by increasing or lowering the voltage on the Control pin.

Figure 9-19 shows the components specified in Figure 9-18 mounted on a breadboard.

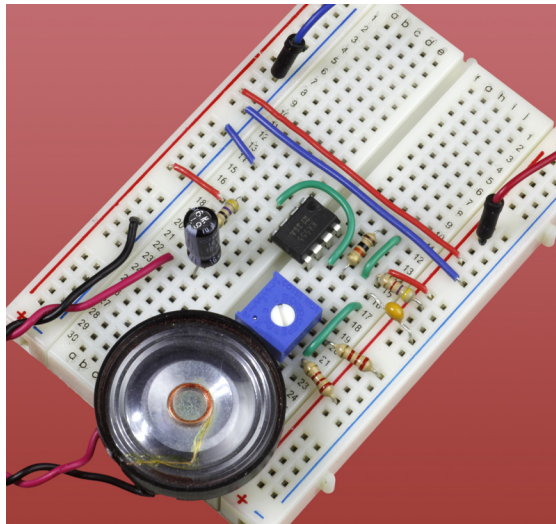


Figure 9-19. The components in the previous schematic are shown here mounted on a breadboard. They will generate an audio output ranging between approximately 425Hz and 1,075Hz. A lower value for the timing capacitor will shift the audio range higher in frequency.

555 Flip-Flop Emulation

The flip-flop inside a 555 timer can be accessed to control the timer's outputs. In Figure 9-20, pushbutton switch S1 applies a negative pulse to the Input pin, creating a high output from the timer, which illuminates LED D1. Normally the pulse length would be limited by the charge time of a capacitor attached to the Threshold pin, but in this circuit there is no capacitor, and the Threshold pin is hard-wired to negative ground. So, it never rises to 2/3 of positive power, and the output from the timer remains high indefinitely.

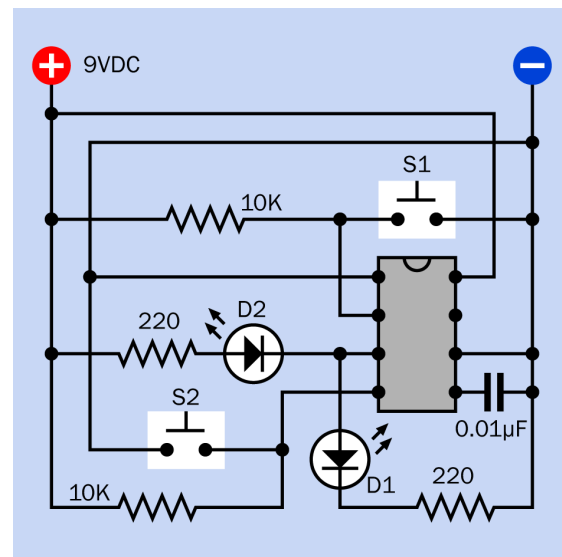


Figure 9-20. A 555 timer can have its timing features disabled so that it functions as a flip-flop.

However, if S2 is pressed, it grounds the Reset pin of the timer, which ends the high output and pulls the Output pin down to a low state. D1 goes out and D2 lights up, as the timer is now sinking current through it. When S2 is released, the timer output remains low and D2 remains illuminated, because the Input pin is held high by a pullup resistor. Therefore, the timer now functions in bi-stable mode, as a flip-flop. While this may be seen as an inappropriate use of the chip, because its full functionality is being disabled, its ability to deliver substantial current and to tolerate a wide range of supply voltages may make it more con-

venient to use than a digital flip-flop. See [Chapter 11](#) for more information about flip-flops.

A 555 timer emulating a flip-flop is shown on a breadboard in [Figure 9-21](#).

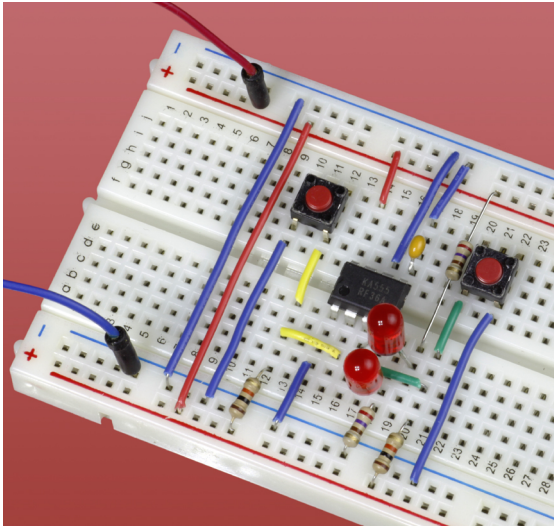


Figure 9-21. The schematic in which a 555 timer acts as a flip-flop is shown here adapted for a breadboard.

555 Hysteresis

The comparators inside a 555 timer enable the chip to produce hysteresis. In [Figure 9-22](#), the Input pin and the Threshold pin are shorted together, and C1, the timing capacitor, is omitted. A 10K potentiometer, wired as a voltage divider, delivers a voltage to the Input pin ranging smoothly from V+ to negative ground. As the input dips below $1/3 V+$, the Output pin goes high, lighting LED D1. Now if the input voltage gradually rises, the output remains high, even as the input rises above $1/3 V+$. The output state is “sticky” because the timer does not end an output pulse until the Threshold pin tells it to, by reaching $2/3 V+$. When this finally occurs, the Output pin goes low, D1 goes out, and D2 comes on, sinking current into the Output pin.

Suppose, now, the input voltage starts to go down again. Once again the output state is “sticky” because it remains low until the Input pin drops below the $1/3$ level. When that happens,

the output finally flips back to a high state, D2 goes out, and D1 comes on.

In the “dead zone” between $1/3$ and $2/3$ of supply voltage, the timer remains in its current mode, waiting for the input to stray outside of those limits. This behavior is known as *hysteresis*, and is of special importance when processing a varying signal, such as the voltage from a temperature sensor, to control an on/off device such as a thermostat. In fact the 10K potentiometer in this demo could be replaced with a thermistor or a phototransistor, wired in series with a resistor to create a voltage divider which will have an input range compatible with the 555 timer. The hysteresis can then be adjusted by varying the supply voltage that powers the timer, as this will change the values of $1/3 V+$ and $2/3 V+$. Alternatively, varying the voltage on the Control pin will affect the hysteresis.

A **comparator** can provide much more versatile control of hysteresis by using positive feedback (see [Chapter 6](#) for additional details). But the 555 timer provides a quick-and-simple substitute, and its greater ability to source or sink current enables it to be connected with a wider range of other components.

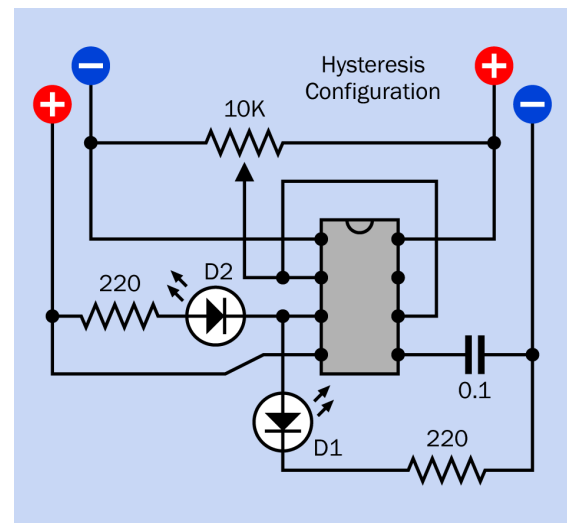


Figure 9-22. A 555 timer wired so that it creates hysteresis, when supplied by a variable input voltage.

555 and Coupling Capacitors

As previously noted, when a basic bipolar 555 timer (and some of its variants) is wired in mono-stable mode, it will retrigger itself indefinitely if its input remains low. One way to avoid this is by using a coupling capacitor. This will pass a transition from high to low, but will then block a steady subsequent voltage. In [Figure 9-23](#), a phototransistor in series with a resistor provides a variable voltage to the noninverting input of a comparator. The reference voltage of the comparator is adjusted with a potentiometer, and resistor R3 provides positive feedback, ensuring that the output from the comparator will be quick and clean. The output from the 555 timer goes through a transistor to the relay, shown at the bottom.

It is important to see the function of the coupling capacitor, C3, with the pullup resistor, R2, which holds the Input pin of the 555 timer high by default. When the output from the comparator drops from high to low, C3 passes this transition to the Input pin of the timer, momentarily overcoming the positive potential, and triggering the timer. After the timer responds, however, C3 blocks any continuation of low voltage from the comparator. Pullup resistor R2 resumes its function of holding the input high, and prevents the timer from being retriggered.

555 Loudspeaker Connection

A small 8-ohm loudspeaker can be driven from the output of a bipolar 555 timer running in astable mode, but should be isolated from it with a 10 μ F to 100 μ F capacitor. A series resistor of 47 Ω (minimum) should be used. See [Figure 9-24](#).

Burst Mode

It is sometimes useful to create a short beep of fixed length in response to a button-press. The beep should terminate even if the button is held down. This “burst mode” can be achieved with the circuit in [Figure 9-25](#), where the button connects power to a bipolar-type 555 timer running in astable mode, and an RC network applies a

decreasing potential to a 47 μ F capacitor, which is wired to the Reset pin of the timer. The resistor in series with the capacitor will vary the length of the beep. When voltage to the pin drops below approximately 0.3V, output from the timer stops and cannot restart until the button is released.

A resistor of greater than 1.5K may not allow the input value at the reset pin to fall below the voltage, which is necessary to enable a reset. If a lower power supply voltage than 9VDC is used, the resistor value should be higher—for example, a 5VDC power supply works well with a 1.5K to 2K resistor.

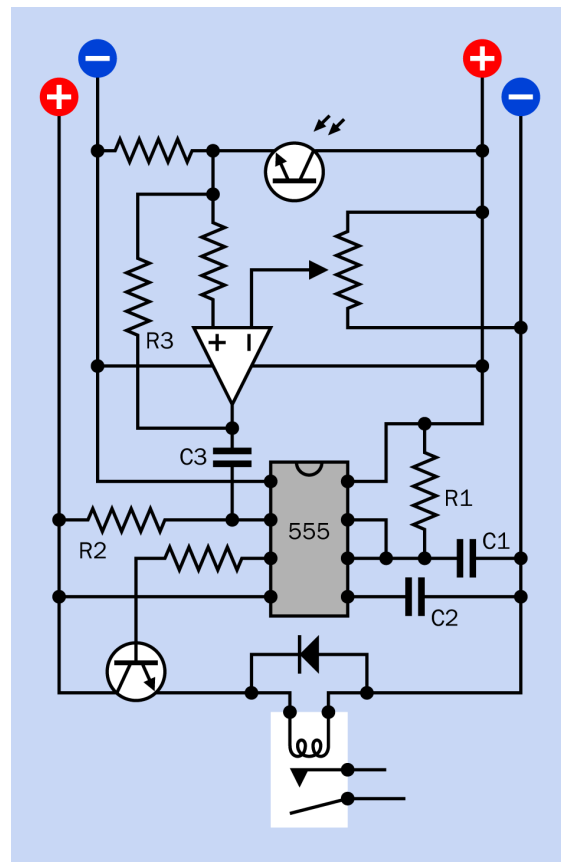


Figure 9-23. A coupling capacitor (C3) is used in this circuit to isolate the 555 timer from a sustained low input from the comparator. The capacitor only passes a transition from high to low. The rest of the time, the pullup resistor (R3) holds the input high.

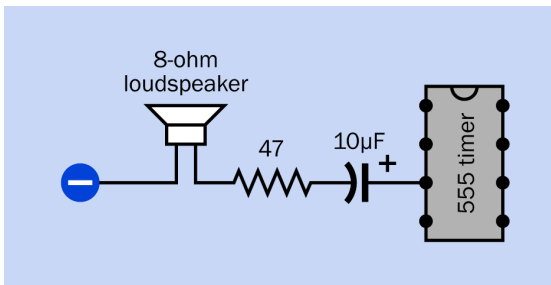


Figure 9-24. A small 8-ohm loudspeaker can be attached through a capacitor and a resistor to the output of a bipolar 555 timer.

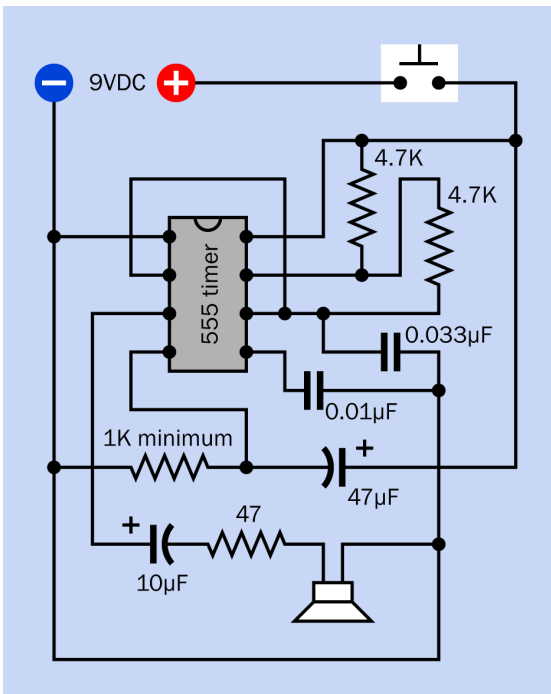


Figure 9-25. An RC circuit, wired to apply a decreasing voltage on the Reset pin of a bipolar 555 timer, will shut off the timer shortly after it is powered up. This can be used to create a fixed-length beep in response to a button press of any duration.

Figure 9-26 shows the components installed on a breadboard.

“You Lose” Game Sound

A timer is a simple, cheap way to create a variety of simple game sounds. The schematic in Figure 9-27 makes a groaning sound as the 100µF

capacitor wired to the Control pin of a bipolar-type 555 gradually charges through the 1K resistor. Note that if a larger resistor is paired with a smaller capacitor, the effect will differ. The 150K resistor is included to discharge the capacitor reasonably quickly in time for the next cycle.

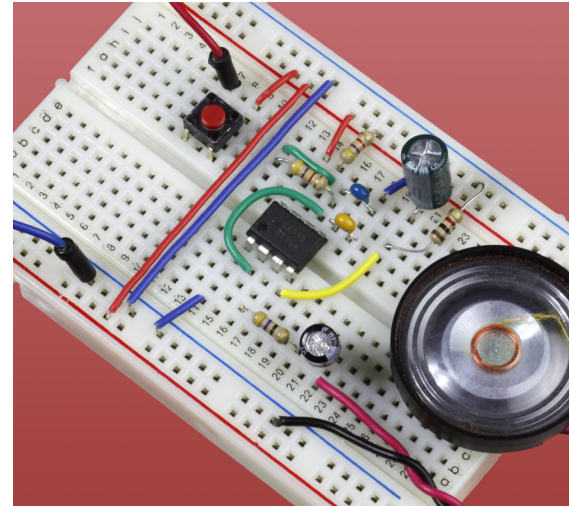


Figure 9-26. The “burst mode” circuit installed on a breadboard with a miniature loudspeaker.

What Can Go Wrong

Dead Timer

Like any chip, the 555 can be damaged by over-voltage, excessive source current or sinking current, static electricity, incorrectly applied polarity of power supply, and other forms of abuse. The TTL version of the timer is fairly robust, but the CMOS type much less so. Check for obvious errors such as lack of supply voltage, incorrect or ambiguous input voltages, and unusual current draw (too high, or none at all, at the V+ pin). Use the meter probes on the actual pins of the chip, in case there is a break in the wiring that feeds them. Because timer chips are cheap, a reserve supply of them should be maintained.

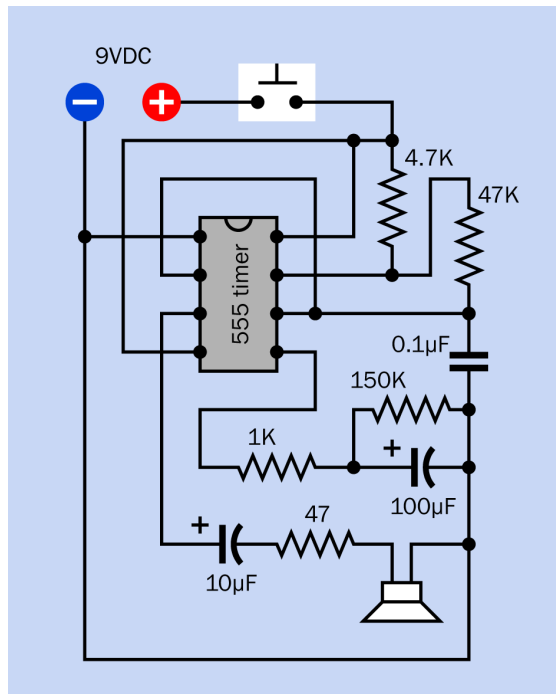


Figure 9-27. An RC circuit, wired to apply an increasingly positive voltage to the Control pin of a bipolar timer running in astable mode, will gradually pull down the frequency at the Output pin, creating a sound that may be useful in simple game applications.

CMOS Confused with Bipolar

The part numbers of some bipolar chips are very similar to those of some CMOS versions, and the chips look physically identical. But the CMOS version is easily overloaded, as it may source only 10mA to 20mA maximum while the TTL version is capable of 200mA. Make sure that your chips are carefully labeled when they are stored.

The Pulse that Never Ends

If a 555 timer responds correctly to a high-to-low transition on the input pin, but the output pulse continues indefinitely, check the voltage on pin 6 to see if the timing capacitor is charging above $2/3$ of $V+$. While a 555 can run from 5VDC, a high-current device on the Output pin can pull down the voltage inside the chip to the point where the capacitor never charges sufficiently to end the cycle.

Also check that the input transition from high to low lasts for a shorter time than the pulse. A persistent low input can retrigger the timer.

Erratic Chip Behavior

Possible causes include:

- Floating pins. The Input pin, in particular, should always be connected with a defined voltage (via a 10K pullup resistor, if necessary), and must not be allowed to float at an indeterminate potential.
- Voltage spikes. A timer can be triggered by transients from other components, especially inductive loads. If the input to a monostable timer dips for even a fraction of a second, the timer will initiate a new cycle. A *protection diode* should be used in conjunction with an inductive load.
- Voltage spikes can also introduce variations in the pulse train from an astable timer.
- TTL versions of the 555 timer will tolerate a wide range of supply voltage, but if a voltage regulator is not used, fluctuations in voltage can have unpredictable consequences.

Interference with Other Components

Because the bipolar version of a 555 timer creates a voltage spike when its output changes state, it can interfere with the normal function of other components, especially CMOS chips. A 0.1µF bypass capacitor can be applied between the timer's $V+$ pin and ground.

Erratic Behavior of Output Devices

If a 555 timer powers an output device such as a relay, and the relay is not opening or closing in a reliable manner, first check that it is receiving sufficient voltage. If the 555 timer is powered with 5VDC, its output will be only around 4VDC.

This problem can be avoided by using the output from the timer to control the voltage on the base

of a transistor which switches a separate source of power to the relay coil.

Fatal Damage Caused by Inductive Loads

While it is possible to drive an inductive load such as a small motor or relay directly from a TTL 555 timer, two precautions should be taken. First, the motor or the coil of the relay should have a clamping diode added around it, as is standard practice. Second, because the output of the timer is capable of sinking current as well as sourcing current, it can be protected from sinking back-EMF by inserting a diode in series with the load. This is illustrated in [Figure 9-28](#).

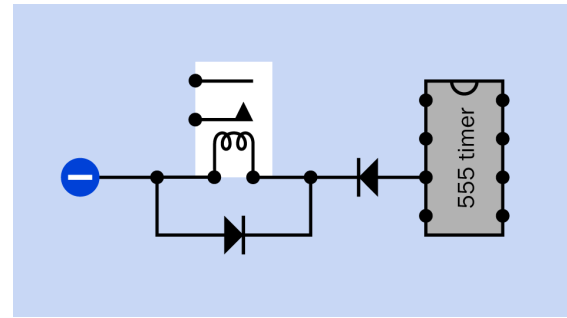


Figure 9-28. In addition to a standard protection diode clamped around an inductive load such as a relay coil, the 555 timer can be protected from back-EMF by adding a diode in series. The series diode must of course be rated to carry sufficient current through the coil. When choosing a relay, allowance must be made for the voltage drop that will be imposed by the series diode.

logic gate

10

Only basic **logic gates** are included in this entry—that is, components that perform a Boolean logic operation on two to eight inputs (or one input, in the case of an inverter) to create a single high or low logical output.

OTHER RELATED COMPONENTS

- **flip-flop** (see [Chapter 11](#))

What It Does

A **logic gate** is a circuit that delivers an output, either high or low, depending on the states of its two inputs, either or both of which can be high or low.

Some gates may have more than two inputs, and an *inverter* only has one input, but the basic gates all conform with the two-input, one-output model. The components that constitute a logic gate are almost always etched into a wafer inside a silicon chip.

In a digital computer, a *high logic state* is traditionally close to 5VDC and represents a value of 1 in *binary arithmetic*, while a *low logic state* is traditionally close to 0VDC and represents a binary 0. Modern devices may use a lower voltage for the high state, but the principle is still the same.

A small network of logic gates can perform binary addition, and all other operations in a digital computer are built upon this foundation.

Origins

The concept of digital logic originated in 1894, when English mathematician George Boole announced his invention of a form of algebra (now referred to as *Boolean algebra*) to analyze com-

binations of two logical states that could be interpreted as “true” and “false.” This concept had few practical applications until the 1930s, when Claude Shannon saw that because a basic switch has two states, Boolean algebra could enable analysis of complex networks of switches that were being used in telephone systems.

Because the state of a switch could also be used to represent the values 0 or 1 in binary arithmetic, and because a transistor could function as a switch, Boolean algebra was implemented in solid-state digital computing equipment.

How It Works

While conventional arithmetic uses *arithmetical operators* to represent procedures such as addition or multiplication, Boolean algebra uses *Boolean operators*. The operators of special interest in digital electronics are named AND, NAND, OR, NOR, XOR, and XNOR.

Although each gate actually contains multiple transistors, it is represented by a single logic symbol, as shown in [Figure 10-1](#). The names of the Boolean operators are customarily printed all in caps. A gate requires a power supply and a connection with negative ground, separate from its inputs, but these connections are omitted from

gate schematics because they are assumed to exist.

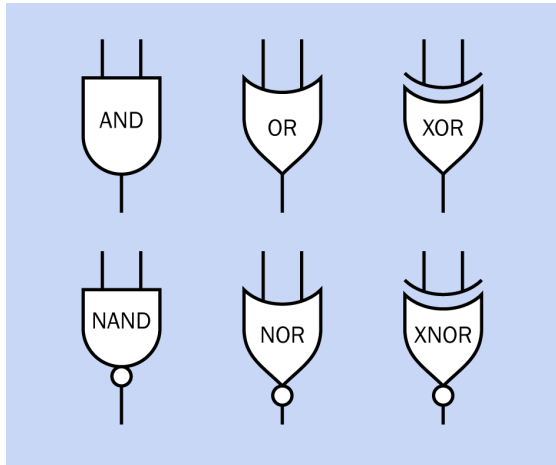


Figure 10-1. Six types of two-input logic gates are used in digital electronics, although the XNOR gate is rare, as it has few applications. The names are customarily printed in uppercase letters.

The functions of the gates with two inputs can be defined in electrical terms. In [Figure 10-2](#), the four possible combinations of inputs are tabulated in the left column, with red indicating a high input and black indicating a low input. The corresponding output from each gate is shown beneath its name. This kind of tabulation is known as a [truth table](#), as it is derived from Boolean algebra which originally concerned itself with “true” and “false” states.

Input states of a two-input gate	Gate Outputs					
	AND	NAND	OR	NOR	XOR	XNOR
● ●	●	●	●	●	●	●
● ●	●	●	●	●	●	●
● ●	●	●	●	●	●	●
● ●	●	●	●	●	●	●

Figure 10-2. The four possible combinations of input states in a 2-input logic gate are shown at left. The corresponding output from each gate is shown beneath its name. Red indicates a high state, while black indicates a low state.

The truth table assumes that [positive logic](#) is being used. Negative logic is very uncommon, but if it were used, the red dots in the truth table would correspond with low inputs and outputs, while the black dots would correspond with high inputs and outputs.

Inversion

The small circles appended to the outputs of NAND, NOR, and XNOR gates mean that the output of each gate is inverted compared with the AND, OR, and XOR gates. This can be seen by inspection of the output states shown in [Figure 10-2](#). The circles are known as [bubbles](#).

Sometimes logic symbols are shown with a bubble applied to one input, as in [Figure 10-3](#). In these cases, the circle indicates that an input must be inverted. More than one gate may be needed to achieve this logic function in an actual circuit. The style is often used to show the inner workings of an IC, using a minimum number of logic symbols.

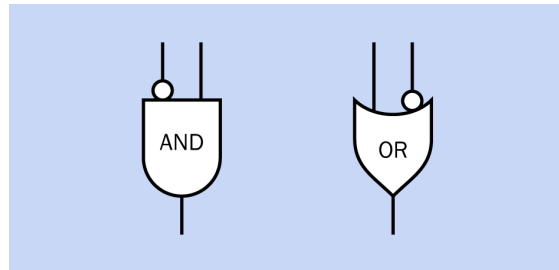


Figure 10-3. The circle in a logic gate symbol indicates that a signal is being inverted. Circles can be inserted at gate inputs, but in a real circuit a separate inverter is likely to be needed to create this effect.

Single-Input Gates

Two gates exist that have a single input and a single output, shown in [Figure 10-4](#). The [buffer](#) should not be confused with the symbol for an **op-amp** or a **comparator**. (Those components always have two inputs.) The output state of a buffer is the same as its input state, but the component may be useful to deliver more current or to isolate one section of a circuit from another.

When a bubble is appended to a buffer, it becomes a NOT gate, more commonly known as an *inverter*. Its function is to create an output state that is opposite to its input state.

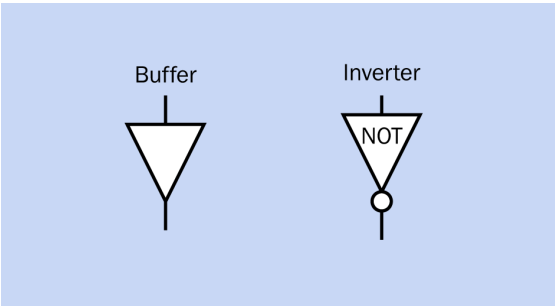


Figure 10-4. The two logic gates that have only one input and one output. Note that in some schematics showing internal logic of ICs, the bubble on an inverter may be found on the input side instead of the output side.

Gates with More than Two Inputs

AND, NAND, OR, and NOR gates can have any number of inputs, as suggested in [Figure 10-5](#), although practical factors usually limit the inputs to a maximum of eight.

Input states of a gate with more than two inputs	Gate Outputs			
	AND	NAND	OR	NOR
All low	●	●	●	●
All high	●	●	●	●
At least 1 low and at least 1 high	●	●	●	●

Figure 10-5. The previous table has been modified to show the outputs from logic gates that have more than two inputs. XOR and XNOR gates are not included in the table, because a strict interpretation of their logic requires that a unique output state exists if one input is high while the other is low.

The rules can be summarized like this:

- Output from an AND gate: Low if any of its inputs is low, high if all of its inputs are high.
- Output from a NAND gate: High if any of its inputs is low, low if all of its inputs are high.

- Output from an OR gate: High if any of its inputs is high, low if all of its inputs are low.
- Output from a NOR gate: Low if any of its inputs is high, high if all of its inputs are low.

In the case of XOR and XNOR gates, their logic requires that a unique output state must exist if one input is high while the other input is low.

In fact, so-called three-input XOR gates do exist, an example being the 74LVC1G386 chip, in which the output is high if all three inputs are high, or if one input is high, but not if two inputs are high or no inputs are high. Further discussion of more-than-two-input XORs is outside the scope of this encyclopedia.

Boolean Notation

For reference, the original written notation for Boolean operators is shown in [Figure 10-6](#). Unfortunately, the notation for these operators was never properly standardized, and in more than one instance, multiple symbols acquired the same meaning. The letters P and Q are often, but not always, used to represent two input states that can be true or false.

- The use of a horizontal line above a symbol, to indicate that its state has been reversed, has carried over to datasheets where this notation can show that an output state from any digital chip is inverted. The line is known as a *bar*.

Arithmetical Operations

Suppose we wish to sum two binary numbers, each containing two digits. There are four digits altogether, and depending on their values, there are 16 different possible addition sums, as shown in [Figure 10-7](#).

If A0 and B0 represent the rightmost digits of the two numbers being added, and S0 is the sum of those two digits, inspection of the figure shows that the sum can be derived using just three rules:

1. If $A0 = 0$ and $B0 = 0$, then $S = 0$.
2. If $A0$ and $B0$ have opposite states, then $S0 = 1$.
3. If $A0 = 1$ and $B0 = 1$, then $S0 = 0$, and carry 1 to the next place left.

If $A0$ and $B0$ are the two inputs to an XOR logic gate, the output of the gate satisfies all three rules, except the need to carry 1 to the next place left. This last function can be satisfied with an AND gate. The function of two gates is known as a *half adder*, and is shown in the top section of Figure 10-8.

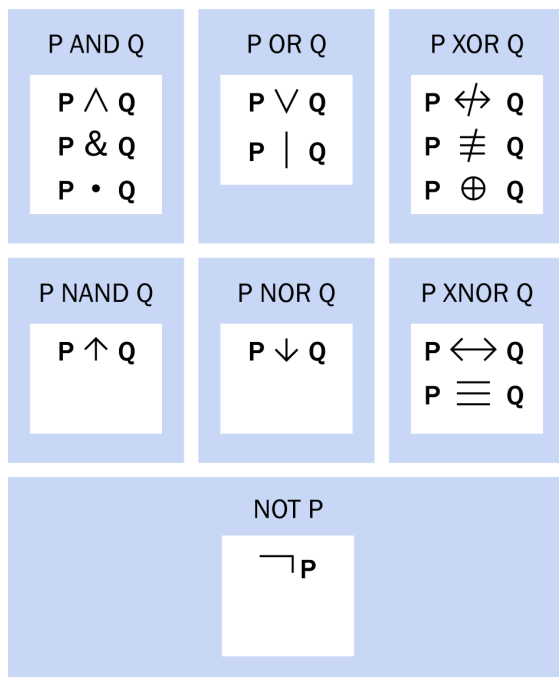


Figure 10-6. Boolean operators as they have been expressed in written notation. Lack of standardization has resulted in more than one symbol representing some of the operators.

When we consider the next pair of binary digits to the left, the situation now becomes more complicated, because we may be carrying 1 into this addition sum from the previous stage, and we still need to be able to carry 1 out (if necessary) to the next stage. An assembly of five logic gates

can deal with this, and their combination is known as a *full adder*. This is shown in the bottom section of Figure 10-8.

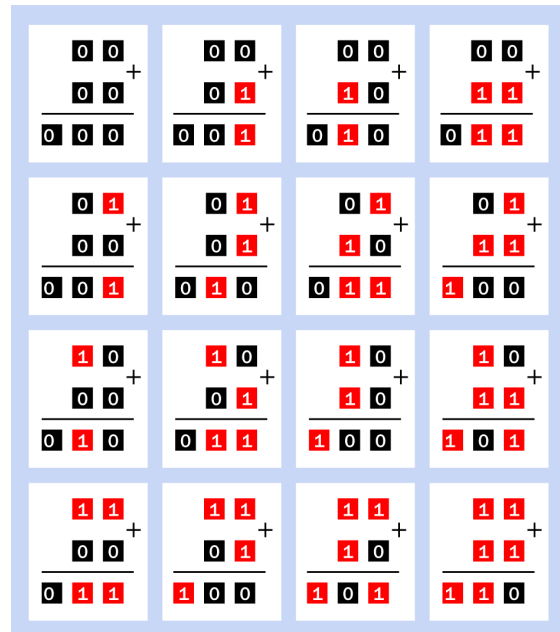


Figure 10-7. Sixteen different addition sums are possible, when summing two binary numbers of two digits each.

The combination of XOR and AND gates shown in Figure 10-8 is not the only one that works to add binary numbers. However, it may be the most intuitively obvious.

Other Operations

Binary arithmetic remains the most important application of logic gates, but individually packaged gates are seldom used for that purpose anymore. They were long since subsumed into large multifunction computing chips.

Single gates still have application in small systems, or to modify the inputs and outputs of microcontrollers, or to convert the output from one complex digital chip to make it compatible with the input of another. This last application is often referred to as *glue logic*.

Applications for single gates are discussed in "How to Use It" on page 103.

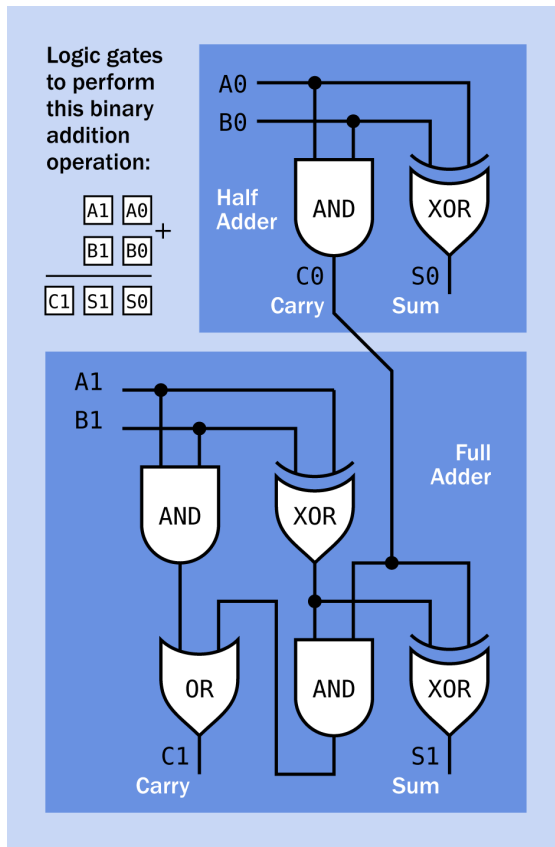


Figure 10-8. Logic gates can be used to add binary numbers, using a high input or output to represent a binary 1 and a low input or output to represent a binary 0. This schematic shows one possible way for gates to add two two-digit binary numbers.

Variants

Chips containing logic gates were introduced in the 1960s. The 7400 NAND chip, from Texas Instruments, was the first of a series that became so influential, the same basic part numbers (with letters added before, after, and among the digits) are still used today. An example of a currently available through-hole 7400 chip is shown in [Figure 10-9](#).

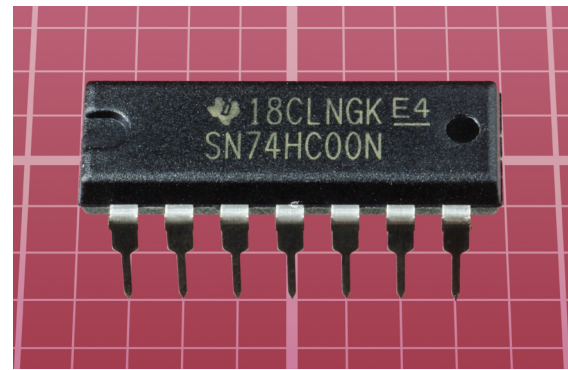


Figure 10-9. A modern version of a 7400 chip containing four NAND gates.

Initially, these chips conformed with a [transistor-transistor logic \(TTL\)](#) standard that had been invented at TRW in 1961 and introduced in commercial products by Sylvania in 1963. It established the now-familiar standard of 5VDC for the power supply. Many logic chips now use other voltages, but the term “high” still means an input or output that is near to the supply voltage, while “low” means an input or output that is near to negative ground. The exact definition of “near” will be found in datasheets for the chips.

The [7400 series](#) was successful partly because it was engineered for compatibility. The output from one gate could be connected directly to the input of another gate, with a few bypass capacitors added on a circuit board to suppress voltage spikes caused by rapid switching. Earlier components had not been so easy to interface with each other. The new standard dominated the industry to the point where dozens of manufacturers started making chips that conformed with it, and a single board could mix-and-match chips from multiple vendors.

Because many logic chips acquired part numbers that began with 74, they are often referred to as the 74xx series, where other digits (sometimes more than two) can be substituted for xx. This avoids ambiguity, as the very first chip in this format was a NAND gate that had 7400 as its actual part number. In the text below, 7400 will refer to

that specific chip, whereas 74xx will refer to the whole series of chips.

RCA introduced a competing family of logic gates in 1968, using CMOS transistors. As each part number began with a 4 and contained four digits, this was referred to as the *4000 series*. The CMOS chips were slower and more expensive, but tolerated a wider range of power supply voltages (3V to 12V, initially). Their biggest advantage was that they used much less current. This was important, as TTL chips created a lot of waste heat. The lower power consumption of CMOS also enabled one chip to control the inputs of many others, which simplified circuit design. This one-to-many relationship is known as *fanout*.

Ultimately, CMOS chips transcended their early limitations. While they were reserved initially for battery-powered devices in which very low power consumption was more important than speed, CMOS is now used almost everywhere, still maintaining its advantage of low current (almost zero, in fact, while a chip is quiescent) while equalling the speed of TTL. However, CMOS logic chips are very often pin-compatible with the old TTL components, and modern CMOS part numbers are often derived from the old 74xx series.

Most CMOS logic chips in the old 4000 series are still available, and may be used in situations where a power supply greater than 5VDC is convenient.

Part Numbers

As the performance of semiconductors gradually improved, successive *families* of logic chips were introduced, identified by one-letter, two-letter, or three-letter acronyms. The acronym was inserted into the part number, so that a 7400 NAND gate in the HC (high-speed CMOS) family became a 74HC00 NAND gate.

Because these chips were available from multiple sources, the part number was also preceded with one or more letters indicating the manufacturer. And because each chip was manufactured in different versions (for example, some complied

with military specifications, while others didn't), letters were also appended to the end of the part number. Today, the appended letters may indicate whether the chip is of the old through-hole format, or conforms with a more recent surface-mount format.

Summing up:

- Prefix: manufacturer ID.
- Numerals, omitting any letters in the middle: Chip functionality.
- Middle letters: Chip family.
- Suffix letters: Package format.

Thus, for example, the actual part number for a 74HC00 NAND chip could be SN74HC00N, where the SN prefix indicates that it is manufactured by Texas Instruments and the N suffix means that it is in plastic dual-inline-pin (DIP) format. (The SN prefix was introduced by Texas Instruments in the earliest days of integrated circuits as an acronym for "semiconductor network," meaning that multiple transistors were "networked" on a wafer of silicon. Other manufacturers used their own schemes for part numbering, and so SN became exclusively identified with Texas Instruments.)

The system of augmenting part numbers has been further extended by inserting 1G, 2G, or 3G immediately after the family identifier, to indicate surface-mount chips that contain one, two, or three logic gates. If the "G" identifier is missing, the chip usually has four logic gates, which was the standard used in the original 74xx series. This rule applies even in surface-mount formats, where the surface-mount pads of four-gate chips have the same functions as the pinouts of the original TTL versions (except in the case of square-format surface-mount chips, which are not discussed here).

When searching a catalog to find a chip by its part number, it helps to remember that searching for a 7400 chip may not find any hits, but searching for a 74HC00 (or any other valid number con-

taining a family identifier) is much more likely to be understood.

A key to understanding part numbers is shown in [Figure 10-10](#). The upper part of the figure is a guide to interpreting numbers on a generic basis, while the lower part interprets the specific part number shown.

Families

As of 2013, the HC family in the 74xx series has become so widely used, it can be considered the default in the traditional DIP 14-pin format. Incremental improvements are still being made, and new families are being introduced, primarily in surface-mount formats which use lower power-supply voltages (down to around 1VDC).

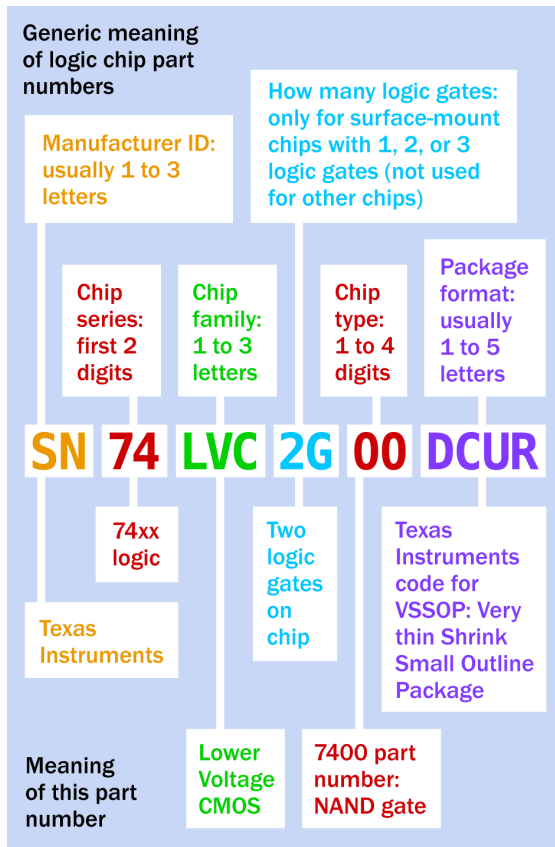


Figure 10-10. How to interpret the segments of a logic chip part number in the 74xx family (in this case, a 7400 NAND gate).

Here is an historical summary of the most important chip families.

- 74xx: Original series of bipolar TTL chips.
- 74Hxx: Bipolar TTL, high speed, about twice as fast as the original 74xx chips, but twice the power consumption.
- 74Lxx: Bipolar TTL, lower power consumption than the original TTL, but also much lower speed.
- 74LSxx: Bipolar TTL, lower power with Schottky input stages, faster than original TTL. Some LS chips are still being manufactured.
- 74ASxx: Bipolar TTL, Advanced Schottky, intended to supersede the 74Lxx.
- 74ALSxx: Bipolar TTL, Advanced Low Power Schottky, intended to supersede the 74LSxx.
- 74Fxx: Bipolar TTL, Faster.
- 74HCxx: CMOS high-speed emulation of 74LSxx.
- 74HCTxx: CMOS but with similar high-state input voltage threshold to bipolar TTL chips, for compatibility.
- 74ACxx: Advanced CMOS.
- 74ACTxx: Advanced CMOS emulation of TTL with similar high-state input voltage threshold to TTL, for compatibility.
- 74AHCxx: Advanced Higher-Speed CMOS, three times as fast as HC.
- 74VHCxx: Very High Speed CMOS.
- 74AUCxx, 74FCxx, 74LCxx, 74LVCxx, 74ALVCxx, 74LVQxx, 74LVVxx: Various specifications, many using power supply voltages of 3.3V or below.

In the 4000 series, an early significant improvement was the 4000B family, which allowed a higher power supply limit (18V instead of 12V) and was much less susceptible to damage by static discharge. The 4000B family almost totally replaced the old 4000 family, and most 4000B

chips are still available, as they are useful in situations where a power supply delivers more than 5VDC.

- When it is referenced casually, the B at the end of a chip number in the 4000 series may be omitted. When the number is listed in a catalog, the B is included.

Chips with 45 as their first two digits were introduced as a new generation, but were not widely adopted. After that, the 4000 series ceased to evolve, and CMOS chips adopted 74xx part numbers, distinguishing themselves by the insertion of letter groups in the center of the number.

To add to the confusion, some 4000 series part numbers were appended to 74xx part numbers, so that, for example, the 74HCT4060 is designed to be compatible with the old 4060B chip.

Family Interoperability

One of the most important issues relating to chip families is their differing specifications for a low-state voltage and a high-state voltage in inputs and outputs.

The original 74xx TTL series, using a 5VDC power supply, used these approximate specifications:

- Output: 74xx voltage representing a low state (at most 0.4V to 0.5V)
- Input: 74xx input voltage interpreted as a low state (maximum 0.8V)
- Output: 74xx voltage representing a high state (at least 2.4V to 2.7V)
- Input: 74xx input voltage interpreted as a high state (minimum 2V)

This provided a safe margin of error of at least 0.4V when chips were communicating with each other.

In the CMOS 4000 family, however, logic chips required a minimum input of 3V to 3.5V to be interpreted as a high state. The minimum acceptable output from a TTL chip was below this level, creating problems if anyone should try to

use the output from a TTL chip to communicate with an input on a CMOS chip.

One solution is to add a 4.7K pullup resistor to the TTL output, guaranteeing that it won't fall too low. But this wastes power, and the need for the resistor is easily forgotten. Another option is to use the HCT or ACT family of CMOS logic. The "T" in these family names indicates that they have been engineered to share the input standards of the old TTL chips. They still deliver the same high output as other types of CMOS, making them seem to be the best possible solution. Unfortunately, it does entail a compromise: the "T" chips are more sensitive to noise, among other factors.

- Ideally, chip families should not be intermingled.

Gates per Chip

Each of the original 74xx chips contained multiple gates within the limits of a uniform 14-pin through-hole format. The gates that were most commonly used had two inputs, and there were four of these gates per chip.

However, the desire for miniaturization, and the use of automatic chip-placement and soldering equipment, made one-gate and two-gate logic chips desirable and practical in surface-mount format. (Three-gate surface-mount chips exist, but are sufficiently unusual that they are not described in this encyclopedia.)

Two Inputs, Single Gate

Where a chip contains just one logic gate, it is almost always a surface-mount component, and the part number has 1G in the middle to indicate "one gate." Pad functions are shown in [Figure 10-11](#). The layout is standardized for all logic gates, with the exception of XNOR gates, which are not manufactured in surface-mount format.

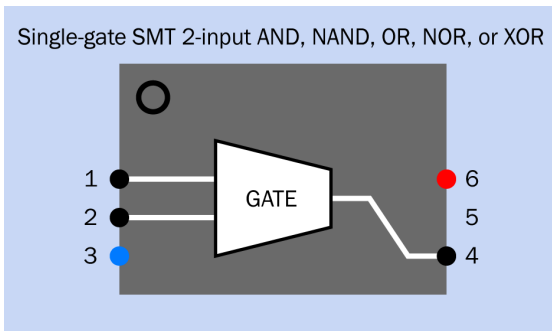


Figure 10-11. Internal configuration and solder-pad functions for a two-input surface-mount single-gate logic chip that can contain an AND, NAND, OR, NOR, or XOR gate. XNOR gates are not manufactured in this format.

In the figure, a gate is shown in generic form, indicating that it may be an AND, NAND, OR, NOR, or XOR gate, depending on the part number of the chip. Inputs are on the left of the gate, while its output is on the right. The chip does not have a solder pad in position 5, but the pad at top right is identified as pin 6 for consistency with the numbering pattern in other surface-mount components where six pads are common.

The generic part numbers for single-gate surface-mount two-input logic chips are shown here, with letter x indicating that letter sequences are likely to be inserted to indicate manufacturer, logic family, and chip format:

- AND gate: x74x1G08x
- OR gate: x74x1G32x
- NAND gate: x74x1G00x
- NOR gate: x74x1G02x
- XOR gate: x74x1G86x

Three Inputs, Single Gate

AND, NAND, OR, and NOR single gates are available with more than two inputs. Their output is determined by rules shown in [Figure 10-5](#). XOR and XNOR gates are not included in the table, because a strict interpretation of their logic requires that a unique output state exists if one input is high while the other is low.

The pad functions for a surface-mount single-gate logic chip with three inputs are shown in [Figure 10-12](#). The generic part numbers for these chips are shown below. Again, each x indicates that letter sequences are likely to be inserted to indicate manufacturer, logic family, and chip format.

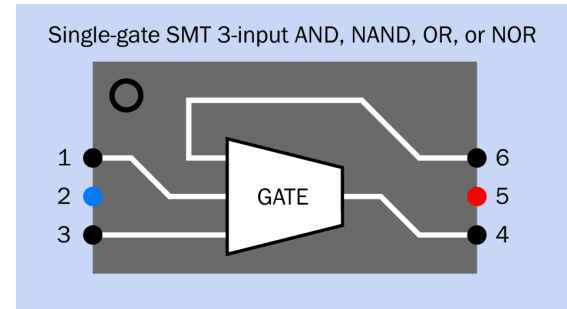


Figure 10-12. Internal configuration and solder-pad functions for a three-input surface-mount single-gate logic chip that can contain an AND, NAND, OR, or NOR gate.

The generic part numbers for single-gate surface-mount three-input logic chips are shown here, with each x indicating that letter sequences are likely to be inserted to indicate manufacturer, logic family, and chip format:

- AND: x74x1G11x
- NAND: x74x1G10x
- OR: x74x1G32x
- NOR: x74x1G27x

Single Gate, Selectable Function

A few surface-mount chips can emulate a variety of two-input gates, by using appropriate external connections. The internal logic of one example, with generic part number x74x1G97x (an actual example would be Texas Instruments SN74LVC1G97), is shown in [Figure 10-13](#). Depending which pin is grounded and which other pins are used as inputs, the chip can emulate all five of the most commonly used gates. To achieve this, however, some inputs have to be inverted.

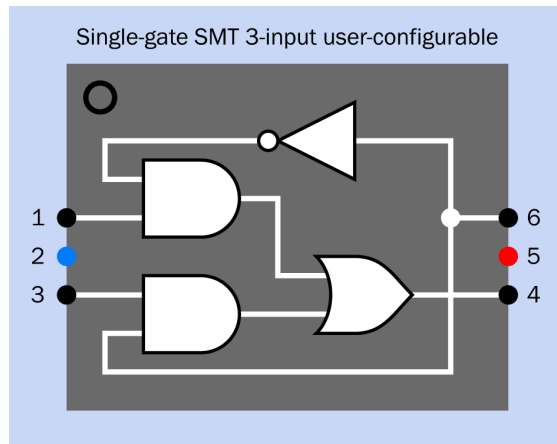


Figure 10-13. Internal configuration for a configurable surface-mount chip that can emulate various two-input logic gates, depending which inputs are used and which are grounded. Some inputs have to be inverted to emulate some gates.

Two Inputs, Dual Gate

Two-input surface-mount AND, NAND, OR, NOR, and XOR gates are available in dual layout (two gates per chip). The internal logic and pad functions are shown in Figure 10-14. The generic part numbers for these chips are shown here. Again, each x indicates that letter sequences are likely to be inserted to indicate manufacturer, logic family, and chip format.

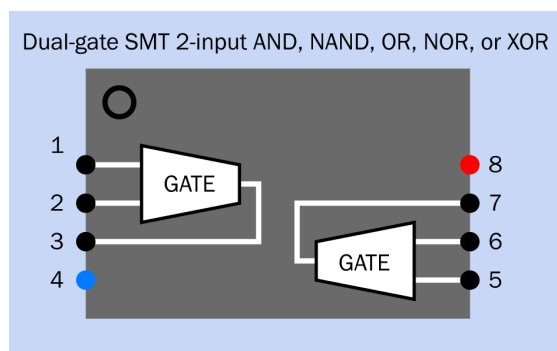


Figure 10-14. Internal configuration and solder-pad functions for a two-input surface-mount dual-gate logic chip that can contain two AND, NAND, OR, NOR, or XOR gates. XNOR chips are not manufactured in this format.

The generic part numbers for dual-gate surface-mount two-input logic chips are shown below, with each x indicating that letter sequences are likely to be inserted to indicate manufacturer, logic family, and chip format:

- AND: x74x2G08x
- NAND: x74x2G00x
- OR: x74x2G32x
- NOR: x74x2G02x
- XOR: x74x2G86x

Original 74xx 14-Pin Format

Each of the original 74xx TTL chips contained multiple gates within the limits of a uniform 14-pin chip format. The available options were, and still are:

- Quad 2-input: Four gates with two inputs each
- Triple 3-input: Three gates with three inputs each
- Dual 4-input: Two gates with four inputs each
- Dual 5-input: Two gates with five inputs each
- Single 8-input: One gate with eight inputs

The five-input chips have become so uncommon that they are not described in this encyclopedia.

Quad Two-Input 74xx Pinouts

14-pin DIP 74xx quad two-input logic chips are available in AND, NAND, NOR, XOR, or XNOR versions, all of which have an internal layout shown in Figure 10-15. The layout is unchanged in surface-mount format. The gates are shown in generic form, as the layout remains the same regardless of which type of gate is in the chip. All the gates in any one chip are of the same type. The four connections leading to a gate are its inputs, while the single connection from a gate is its output.

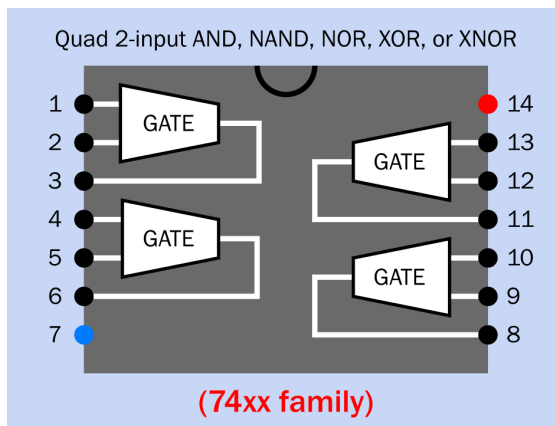


Figure 10-15. In a 14-pin quad two-input 74xx logic chip, the AND, NAND, NOR, XOR, and XNOR versions all share this generic layout.

- The 14-pin quad two-input OR chip has different pinouts from all the other 74xx logic chips. It is shown in Figure 10-16.

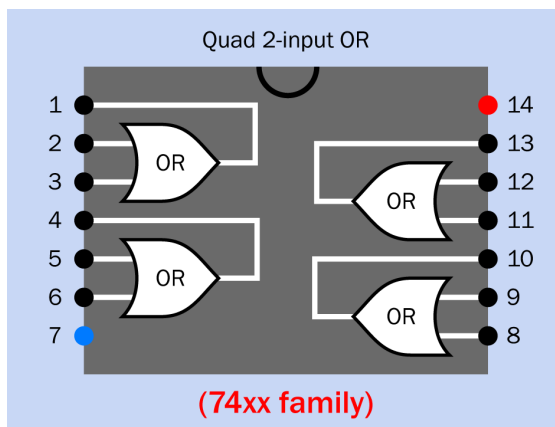


Figure 10-16. In a quad two-input 74xx OR chip, this layout is used, which is different from that used in all the other quad two-input logic gates.

Triple Three-Input 74xx Pinouts

The AND, NAND, and NOR versions of a 14-pin DIP 74xx triple three-input logic chip all have an internal layout shown in Figure 10-17. The layout is unchanged in surface-mount format. The gates are shown in generic form, as the layout remains the same regardless of which type of gate is in the chip. All the gates in any one chip

are of the same type. Three connections leading to a gate are its inputs, while a single connection from a gate is its output.

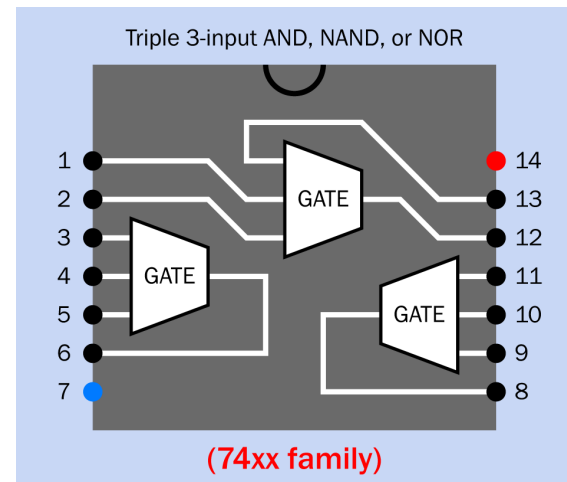


Figure 10-17. In a 14-pin triple three-input 74xx logic chip, the AND, NAND, and NOR versions all share this generic layout.

- The 14-pin triple three-input OR chip has different pinouts from all the other 74xx logic chips. It is shown in Figure 10-18.

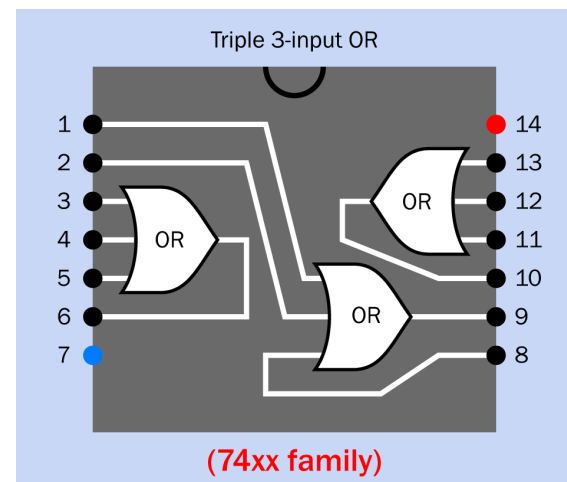


Figure 10-18. In a triple three-input 74xx OR chip, this layout is used, which is different from that used for all the other triple three-input logic gates.

Dual Four-Input 74xx Pinouts

A 14-pin DIP 74xx dual four-input logic chip contains two four-input gates. The AND, NAND, and NOR versions all have an internal layout shown in Figure 10-19. The layout is unchanged in surface-mount format. The gates are shown in generic form, as the layout remains the same regardless of which type of gate is in the chip. All the gates in any one chip are of the same type.

- There is no OR chip of the 14-pin dual four-input type.

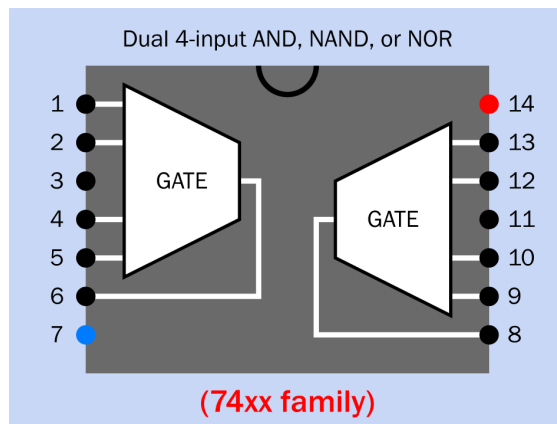


Figure 10-19. In a 14-pin dual four-input 74xx logic chip, the AND, NAND, and NOR versions all share this generic layout. There is no 74xx OR chip with four inputs per gate.

Single Eight-Input 74xx Pinouts

A 14-pin DIP 74xx single eight-input NAND chip contains one eight-input gate, as shown in Figure 10-20. The layout is unchanged in surface-mount format.

- There is no AND chip of the 14-pin single eight-input type.

A 14-pin eight-input logic chip in the 74xx series, able to function as both an OR and a NOR, is shown in Figure 10-21. The output from the NOR gate is connected with pin 13, but also passes through an inverter to create an OR output at pin 1. (Because a NOR gate is equivalent to an

inverted-OR, when its output is inverted again, it returns to being an OR.)

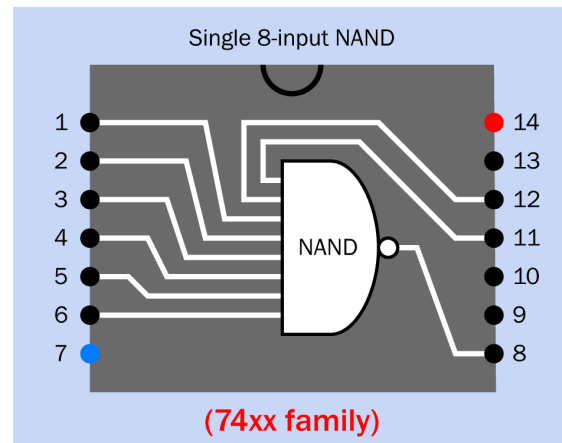


Figure 10-20. The internal layout of single eight-input NAND chip in the 14-pin 74xx series. There is no 74xx AND chip with eight inputs per gate.

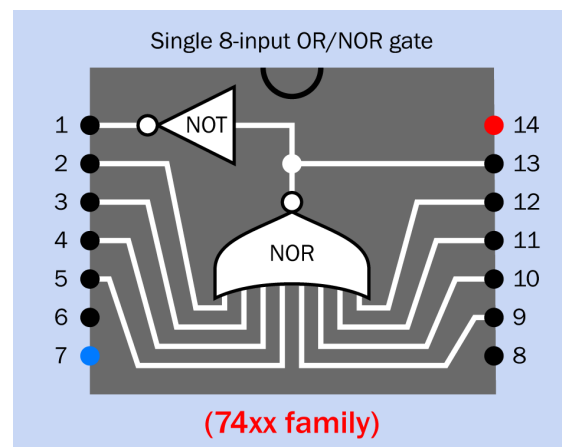


Figure 10-21. The internal layout of single eight-input OR/NOR chip in the 14-pin 74xx series. Pin 13 has the NOR output, while pin 1 has the OR output.

The following list shows the generic part numbers for DIP and surface-mount versions of 14-pin logic chips in the 74xx series that have two or more inputs per gate. As before, an x indicates that letter sequences are likely to be inserted to indicate manufacturer, logic family, and chip format.

- Quad 2-input AND: x74x08x
- Quad 2-input NAND: x74x00x
- Quad 2-input OR: x74x32x
- Quad 2-input NOR: x74x02x
- Quad 2-input XOR: x74x86x
- Quad 2-input XNOR: x74x266x
- Triple 3-input AND: x74x11x
- Triple 3-input NAND: x74x10x
- Triple 3-input OR: x74x4075x
- Triple 3-input NOR: x74x27x
- Dual 4-input AND: x74x21x
- Dual 4-input NAND: x74x20x
- Dual 4-input NOR: x74x4002x
- Single 8-input NAND: x74x30x
- Single 8-input OR/NOR: x74x4078x

74xx Inverters

Single, dual, and triple inverter packages in the 74xx series are available in surface-mount format only. Their internal arrangement is shown in Figures 10-22, 10-23, and 10-24.

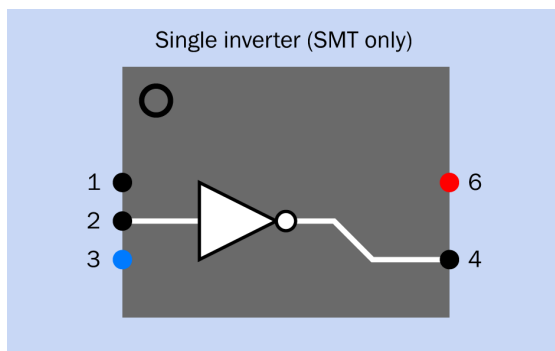


Figure 10-22. The internal layout of a 74xx series logic chip containing one inverter. This is available in surface-mount format only. Pin 5 is absent. Pin 1 is not connected.

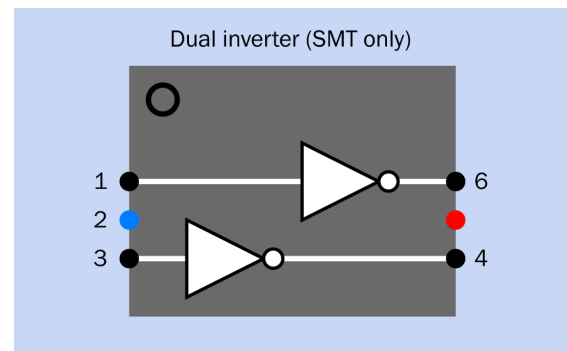


Figure 10-23. The internal layout of a 74xx series logic chip containing two inverters. This is available in surface-mount format only.

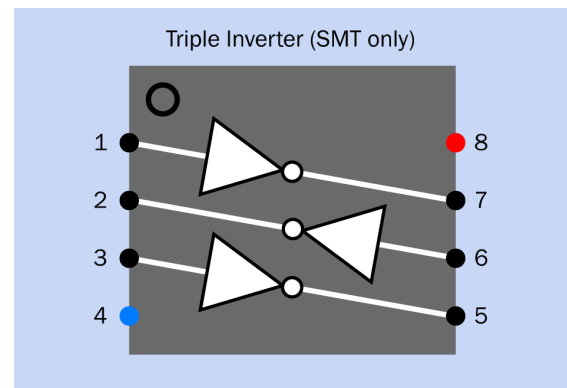


Figure 10-24. The internal layout of a 74xx series logic chip containing three inverters. This is available in surface-mount format only.

In the 14-pin format, a hex inverter chip (containing six inverters) is available, as shown in Figure 10-25. The layout is the same for DIP and surface-mount formats.

Generic part numbers for inverter chips are as follows:

- Single inverter: x74x1G04x
- Dual inverter: x74x2G04x
- Triple inverter: x74x3G14x
- Hex inverter: x74x04x

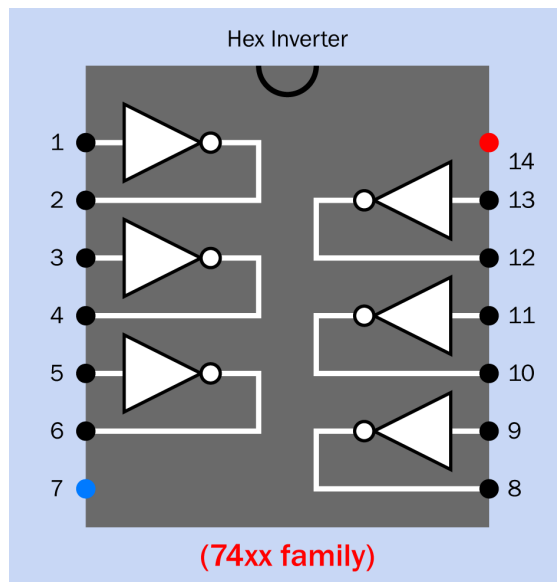


Figure 10-25. The internal layout of a 14-pin 74xx hex inverter logic chip, containing six inverters. This layout is the same for DIP and surface-mount versions.

Additional Variations

Some chips in the 74xx series (both DIP and surface mount versions) have variants with open drain or open collector outputs, while others have inputs that are configured as Schmitt triggers. These variants will be found as hits when searching supplier websites for logic chips by gate name and number of inputs.

Pinouts in the Original 4000 Series

Each of the original 4000 CMOS chips contained multiple gates within the limits of a uniform 14-pin chip format. The available options were, and still are:

- Quad 2-input: Four gates of two inputs each
- Triple 3-input: Three gates of three inputs each
- Dual 4-input: Two gates of four inputs each
- Single 8-input: One gate of eight inputs

In the 4000 family, 14-pin quad two-input logic chips are available in AND, OR, NAND, NOR, XOR, and XNOR versions, all of which have an internal layout shown in Figure 10-26. The gates are shown in generic form, as the layout remains the same regardless of which type of gate is in the chip. All the gates in any one chip are of the same type. The four connections leading to a gate are its inputs, while the single connection from a gate is its output.

or XNOR versions, all of which have an internal layout shown in Figure 10-26. The gates are shown in generic form, as the layout remains the same regardless of which type of gate is in the chip. All the gates in any one chip are of the same type. The four connections leading to a gate are its inputs, while the single connection from a gate is its output.

Unlike the 74xx family, the quad two-input OR chip in the 4000 family has the same pinouts as the other types of quad two-input logic chips.

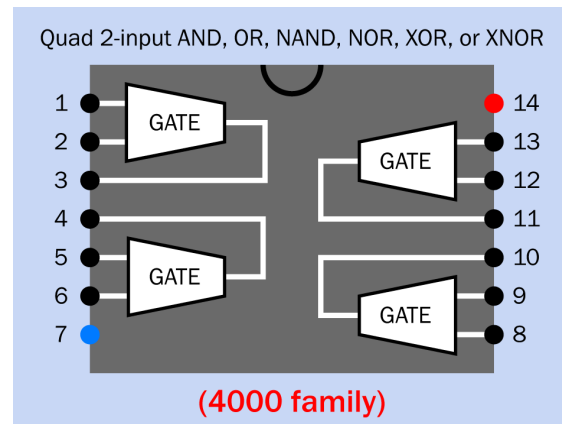


Figure 10-26. In the 4000 family, the AND, OR, NAND, NOR, XOR, and XNOR versions of a quad two-input logic chip all share this generic layout.

In the 4000 family, a 14-pin triple three-input logic chip contains three three-input gates. The AND, OR, NAND, and NOR versions all have an internal layout shown in Figure 10-27. The gates are shown in generic form, as the layout remains the same regardless of which type of gate is in the chip. All the gates in any one chip are of the same type. The three connections leading to a gate are its inputs, while a single connection from a gate is its output.

Unlike the 74xx family, the triple three-input OR chip in the 4000 family has the same pinouts as the other types of triple three-input logic chips.

In the 4000 family, a 14-pin dual four-input logic chip contains two four-input gates. The AND, NAND, OR, and NOR versions all have an internal

layout shown in [Figure 10-28](#). The gates are shown in generic form, as the layout remains the same regardless of which type of gate is in the chip. All the gates in any one chip are of the same type. Each pair of connections leading to a gate are its inputs, while the single connection from a gate is its output.

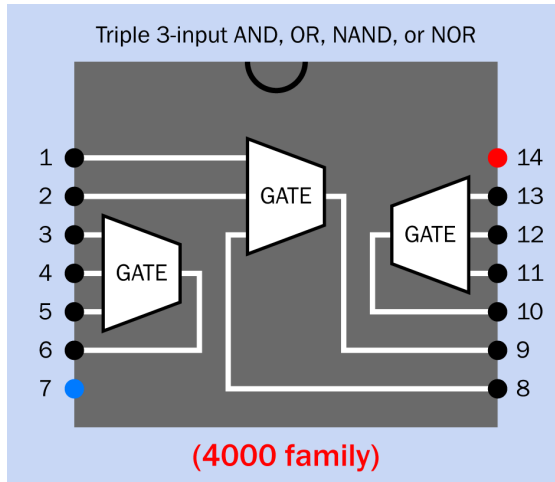


Figure 10-27. In the 4000 family, the AND, OR, NAND, and NOR versions of a triple three-input logic chip all share this generic layout.

Note that the 4000 family does have a dual four-input OR chip, whereas the 74xx family does not.

In the 4000 family, a 14-pin eight-input logic chip with AND and NAND outputs is available, as shown in [Figure 10-29](#).

The following list shows the generic part numbers for 14-pin logic chips in the 4000 family that have two or more inputs per gate (in actual part numbers, letters will be substituted where an x appears):

- Quad 2-input AND: x4081x
- Quad 2-input NAND: x4011x
- Quad 2-input OR: x4071x
- Quad 2-input NOR: x4001x
- Quad 2-input XOR: x4070x
- Quad 2-input XNOR: x4077x

- Triple 3-input AND: x4073x
- Triple 3-input NAND: x4023x
- Triple 3-input OR: x4075x
- Triple 3-input NOR: x4025x
- Dual 4-input AND: x4082x
- Dual 4-input NAND: x4012x
- Dual 4-input OR: x4072x
- Dual 4-input NOR: x4002x
- Single 8-input AND/NAND: x4068x

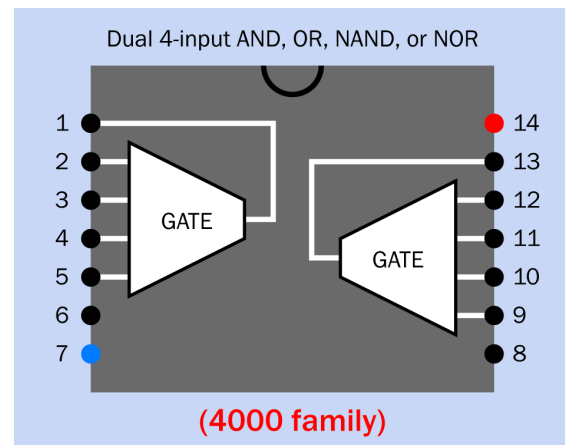


Figure 10-28. In the 4000 family, the AND, OR, NAND, and NOR versions of a dual four-input logic chip all share this generic layout.

4000 Series Inverters

In the 4000 family, the 4069B is a 14-pin hex inverter chip (containing six inverters), as shown in [Figure 10-30](#). This has the same pinouts as the x74x04x chip.

How to Use It

Which Family

In DIP format, the HC family has existed for more than 30 years, and has become established as a widely used default choice.

In surface-mount formats, the choice of family will largely be determined by the choice of supply voltage.

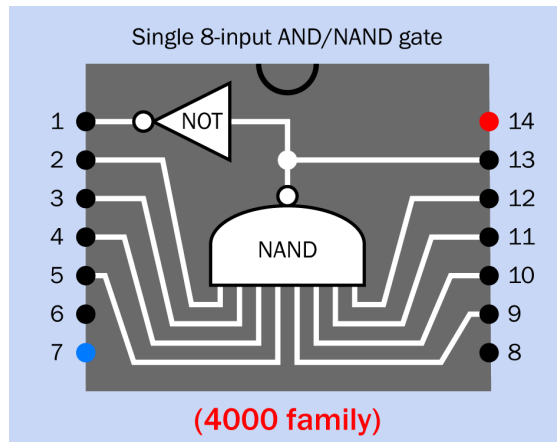


Figure 10-29. In the 4000 family, a single eight-input AND/NAND chip has this internal layout. The inverted output from the NAND gate becomes an AND output from pin 1 of the chip.

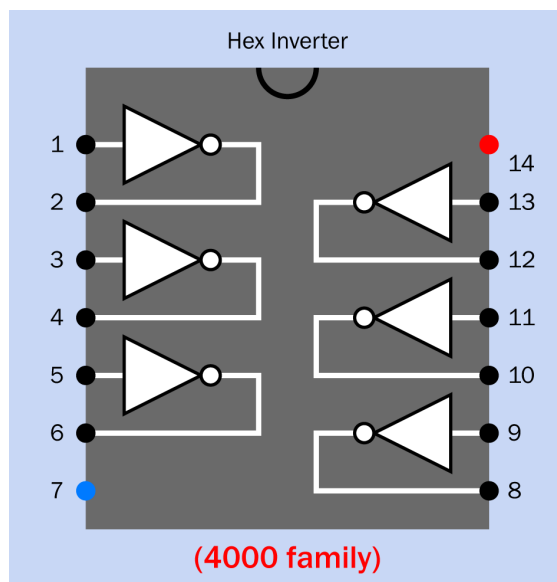


Figure 10-30. The internal layout of a 14-pin 4069B hex inverter logic chip, containing six inverters. This layout is the same as for the x74x04x chip.

Although the 4000 series is now more than 40 years old, it may still be useful where a 5VDC power supply is not required for other reasons in a circuit and would be added purely to power a 74xx series logic gate or other digital chip. If a

circuit contains a 9VDC or 12VDC relay, for instance, a Darlington pair may be used with that voltage to drive the relay, and an old-school 4000 series logic chip could share the same supply. The relay coil would need a clamping diode to protect the logic chip from transients.

Applications

The output from a logic chip may be used as an input for a microcontroller, to enable multiple inputs to share one pin. An eight-input NAND gate, for instance, could combine the inputs from eight normally on motion sensors. If just one sensor responds to an intrusion, the gate output would change from high to low.

Logic gates may be useful in any simple device that has to respond to a single, specific combination of inputs. A digital combination lock is one example; games of chance are another. Most simple dice simulations use logic gates to convert the output from a counter to drive a dice-pattern of LEDs.

A logic gate may be used as an interface between an electromechanical **switch** and a circuit containing digital chips. A 10K pullup or pulldown resistor prevents the gate input from floating when the switch is open. A buffer can be used for this purpose, or an inverter, or any “spare” gate on a logic chip that is already in the circuit. One input of the chip can be tied to the positive power supply or negative ground, to create an appropriate input from the chip when the switch, attached to the other input, is opened or closed.

A jam-type **flip-flop** can be used to *debounce* a switch input. See [Chapter 11](#) for details. If two NOR or two NAND gates are unused in a circuit, they can form a flip-flop.

In the original CMOS 4000 family, a positive output may be capable of driving an LED if the current does not exceed 5mA with a power supply of 5VDC or 10mA with a power supply of 10VDC. Note that the output voltage will be pulled down significantly by these loads. In the 74HCxx family, chips can source or sink as much as 20mA, but

here again the output voltage will be pulled down. Note that the total limit for all outputs from a 74HCxx chip is around 70mA.

The output from a logic chip can be passed through a buffer such as the 7407, which has an open-collector output capable of sinking as much as 200mA. This enables direct drive of modest loads, so long as they are not inductive.

Solid-state relays and **optocouplers** can be driven directly from logic chips, as they draw very little current. A solid-state relay can switch 50A or more.

What Can Go Wrong

Two problems are common when using CMOS digital chips: damage from *static electricity*, and erratic behavior caused by *floating pins*.

Static

The early 4000 series CMOS chips were especially vulnerable, but more recent CMOS designs generally include diodes at the inputs, which reduces the risk. Still, logic chips should be protected by inserting them into anti-static foam or enclosing them in conductive wrappers until they are installed in a board. While handling chips, it is good practice to be grounded, ideally using a wrist-mounted ground wire.

Floating Pins

Any pin which is unconnected in a logic chip is considered to be “floating,” and can pick up signals by capacitive coupling, possibly disrupting the behavior of the chip and also causing power consumption, as the ambiguous pin state will tend to prevent that gate in the chip from entering quiescent mode.

Generally speaking, input pins in a TTL logic chip that are not being used for any purpose should be tied to the positive voltage supply, while unused CMOS pins should be tied to negative ground.

Family Incompatibilities

As previously noted, older TTL logic chips may deliver a “high” output voltage that is lower than the minimum expected by newer CMOS logic chips. The best option is not to mix families, but if chips are stored carelessly, some intermingling can occur. Part numbers should be checked if one chip appears to be ignoring output from another.

Overloaded Outputs

If a circuit calls for a logic chip with an open-collector output, and a regular chip is used by mistake, it will almost certainly be damaged.

Output Pulled Down

If the output from one logic chip is connected with the input of another logic chip, and if the output from the first chip is also connected to an **LED**, the LED may pull down the output voltage so that the second chip will not recognize it as a high state. As a general rule, a logic output can drive an LED, or can drive another logic chip, but not both. Very-low-current LEDs, which draw as little as 2mA, may be acceptable.

Incorrect Polarity and Voltages

Logic chips can be knocked out by applying incorrect polarity, or voltage to the wrong pin, or the wrong voltage. Modern logic chips tolerate a very limited voltage range, and a 74xx series chip will be irrevocably damaged if it is used where a 4000 series chip was specified for a power supply higher than 6VDC.

If a chip is inserted upside-down, it will probably be damaged when voltage is applied.

Bent Pins

Like all through-hole chips, DIP logic chips can be inserted accidentally with one or more pins bent underneath the chip. This error is very easy to miss. The bent pins will not make contact with any socket that is used, and the chip will behave unpredictably. Check for proper pin insertion with a magnifying glass if necessary.

Unclean Input

Logic chips expect a clean input without voltage spikes. A 555 timer of TTL type generates spikes in its output which can be misinterpreted as multiple pulses by the input of a logic chip. A CMOS-type 555 timer is more suitable for connection with logic chips.

If a pushbutton, rotational encoder, or electro-mechanical switch provides a high or low input, the input must be debounced. In hardware, this is traditionally done with a **flip-flop**. It can also be done with code in a microcontroller.

Analog Input

The input of a logic chip can be connected directly with a thermistor, phototransistor, or sim-

ilar analog component, but only if there is some certainty that the voltage at the input pin will remain within the range that is acceptable to the chip. In the case of a phototransistor, for example, it should be exposed to a limited, known range of light intensity.

In general, it is best to avoid applying intermediate-voltage signals to a digital logic input, as they can create unpredictable output, or output of an intermediate voltage. A **comparator** can be placed between the analog source and the digital logic chip, or a logic chip with a Schmitt-trigger input can be used.

flip-flop

The term **flip-flop** is sometimes printed with a space instead of a hyphen, but the hyphenated form seems to predominate in the United States. Therefore, the hyphen is included here. The term *flipflop* (with neither a hyphen nor a space) is sometimes seen, but is unusual. The acronym *FF* is confined mostly to logic diagrams or schematics.

The term *latch* is sometimes used interchangeably with *flip-flop* but is assumed here to describe a minimal asynchronous circuit that responds immediately and transparently to an input. A flip-flop can function as a latch and also as a synchronous device which is opaque, meaning that the input does not flow directly through to the output.

OTHER RELATED COMPONENTS

- **counter** (see [Chapter 13](#))
- **shift register** (see [Chapter 12](#))

What It Does

Transistors enable logic gates; logic gates enable flip-flops; and flip-flops enable many mathematical, storage, and retrieval functions in digital computing. Most flip-flops today are embedded in much larger integrated circuits that have complex functions. However, they are still available as separate components in chip form, and will be discussed on that basis here.

A flip-flop is the smallest possible unit of memory. It can store a single bit of data, represented by either a high or low *logic state*. (A full explanation of logic states is included in the **logic gate** entry. See [Chapter 10](#).) Flip-flops are especially useful in **counters**, **shift registers**, and *serial-to-parallel converters*.

A flip-flop circuit can be classified as a form of *bistable multivibrator*, as each of its outputs is stable in one of two states until an external trigger stimulates it to “flip” from one state and “flop” into the other. (For a comprehensive discussion

of monostable and astable multivibrators, see the **timer** entry in [Chapter 9](#).)

An *asynchronous* flip-flop will respond immediately to a change of input, and can be used for applications such as debouncing the signal from an electromechanical **switch** or building a *ripple counter*. More often, a flip-flop is *synchronous*, meaning that a change in input state will be unrecognized until it is enabled by a low-to-high or high-to-low transition in a stream of pulses from an external *clock*.

How It Works

Every flip-flop has two outputs, each of which may have a high or low state. When the flip-flop is functioning normally, the outputs will be in opposite logical states, one being high while the other is low. These outputs are typically identified as Q and NOT-Q (the latter term meaning a letter Q with bar printed above it, sometimes referred to verbally as “Q-bar”). In datasheets and other documents where a bar symbol cannot be

represented easily above a letter, the NOT-Q output may be represented as letter Q with an apostrophe after it, as in Q'.

Almost always, in a schematic diagram, a flip-flop is represented by a simple rectangle, with inputs and outputs identified by letters and other marks. Because a description of the inner workings is necessary before the different types of flip-flop can be understood, schematic symbols for various flip-flops will not be introduced until “Variants” on page 116.

The simplest flip-flop contains two logic gates whose function can be most easily understood if the inputs are controlled by a SPDT switch. It can be created from two NAND gates or two NOR gates, as described next. This type of component can be described as:

- **asynchronous:** Will accept data on an impromptu basis, as it is not synchronized with a clock.
- **jam-type:** Colloquial equivalent of **asynchronous**. The input is jammed in at any time, forcing an immediate change of output.
- **transparent:** The input state flows straight through to the output.

NAND-Based SR Flip-Flop

Figure 11-1 shows two NAND gates attached to a SPDT switch, with two pullup resistors. When either of the NANDs has a floating input from the switch, the pullup resistor attached to that input will maintain it in a high state. The data inputs for the NAND gates are labeled S and R, meaning Set and Reset, giving this component its name as an **SR** flip-flop:

- In a NAND-based SR flip-flop, a low state is considered an active logic input, as indicated by the bar placed above each letter.
- A high state is considered an inactive logic output.

The schematic style in this figure, with diagonally crossing conductors, is universally used and easily recognizable. The equivalent schematic in Figure 11-2, which might be created by circuit-drawing software, has the same functionality but would not be immediately recognizable as a flip-flop. The “classic” crossed-conductor representation is preferable.

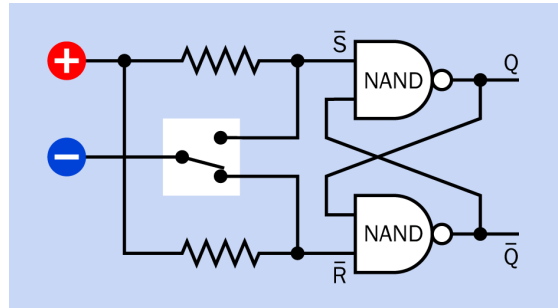


Figure 11-1. The schematic for a simple NAND-based SR-type flip-flop, with a switch and pullup resistors added to control two inputs.

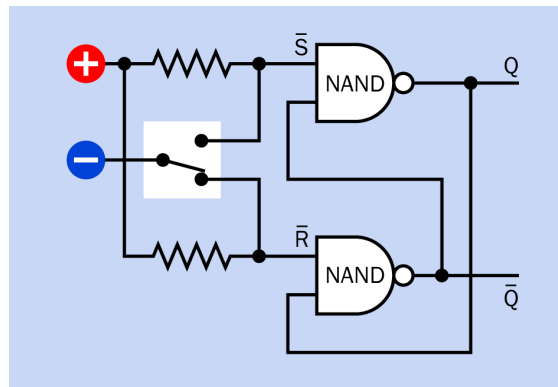


Figure 11-2. An alternative component layout for an SR flip-flop, functionally identical to the previous schematic, but not so easily recognizable. The layout with a pair of diagonally crossing conductors has become so ubiquitous, it should be considered to be a standard.

The first step toward understanding the behavior of flip-flops is to recall the relationship between the two inputs and the output of NAND or NOR gates. This is shown in Figure 11-3, where red indicates a high logic state and black indicates a low logic state.

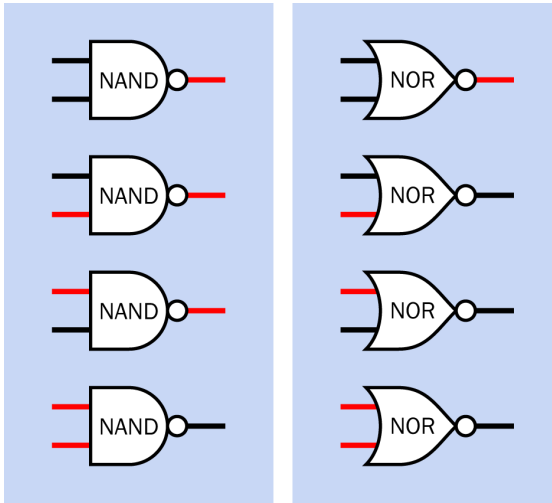


Figure 11-3. The four possible input combinations for a NAND gate and a NOR gate, with the corresponding logical outputs. A flip-flop can be built from two NANDs or two NORs.

The behavior of a NAND gate can be summarized:

- Both inputs high: Output low.
- Other input combinations: Output high.

Figure 11-4 shows a series of snapshots of the SR flip-flop circuit as the switch moves from one position, through an intermediate state where it makes no connection, to the other position. Remember that in this circuit, the active logical input state is low, and the active logical output state is high.

In the top panel, the pullup resistor of the lower NAND is overwhelmed by the direct connection to negative ground, which holds the R input in a low state. The other input of this gate is irrelevant, because the output from a NAND will be high if either of its inputs is low. So, the lower NAND has a high output, which feeds back to the secondary input of the upper NAND gate. The S input of this gate is high, because of the pullup resistor. Because both inputs of this gate are now high, its output is low, which feeds back to the lower gate. The lower gate doesn't change its

output, because either of its low inputs is enough to keep its output high. So, the circuit is in equilibrium. A high state on the NOT-Q output is known as the **Reset** state for a NAND-based flip-flop.

The second panel shows what happens if the switch now moves up into a neutral, disconnected position. The R input of the lower NAND now becomes high, because of the pullup resistor. But this NAND still has one low input, supplied by the output of the upper gate, so its output remains high, and the circuit is still in equilibrium. This is known as the **Hold** state for the NAND circuit.

Suppose the switch bounces to and fro between the states shown in the first two panels. The output from the circuit won't change. This shows that the circuit provides a method for eliminating **switch bounce**—the very fast, momentary spikes that occur when the mechanical contacts of a switch open and close.

The third panel shows what happens if the switch now moves to its upper position. The top input of the upper NAND gate is now pulled low. Consequently, its output goes high. This feeds back to the lower gate. Its other input is high because of the pullup resistor. With both of its inputs high, its output goes low. The gate outputs have flipped and swapped values. A high state on the Q output is known as the **Set** state for a NAND-based flip-flop.

The circuit still remains in equilibrium even if the switch returns to its central, disconnected position shown in the bottom panel. Therefore, the debouncing capability of the circuit works equally well for both positions of the switch.

NOR-Based SR Flip-Flop

Figure 11-5 shows a similar circuit using two NOR gates attached to an SPDT switch. Because the NOR gates function differently, this circuit uses active-high input logic, and pulldown resistors are needed instead of pullup resistors. The output from the circuit still uses active-high logic, and is identical with the NAND-based circuit in

this respect, although the relative positions of the Q and NOT-Q outputs have been swapped.

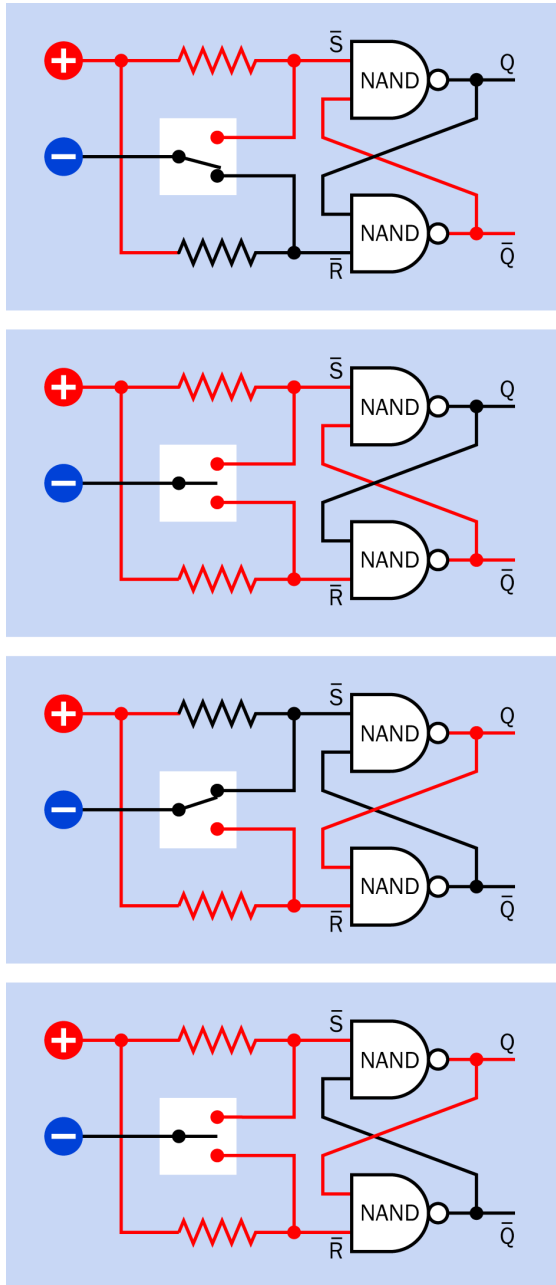


Figure 11-4. Four snapshots of transitions in a NAND-based SR flip-flop as the switch moves down and up through an intermediate no-connection zone. See text for details.

- In a NOR-based SR flip-flop, a high state is considered an active logic input, as indicated by the absence of a bar placed above the letters S and R.
- A high state is considered an active logic output.

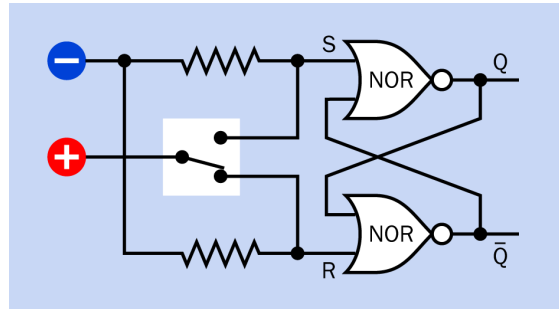


Figure 11-5. The schematic for a simple SR flip-flop using NOR gates instead of NANDs.

In the case of the NOR gate:

- Both inputs low: Output high.
- Other input combinations: Output low.

Figure 11-6 shows a series of snapshots as the switch moves from one position to the other, through intermediate states where it makes no connection. Remember that in this circuit, the active logical state is high at both the inputs and the outputs.

In this circuit, as in the previous circuit using NAND gates, it will ignore switch bounce, allowing the gate outputs to remain stable.

Forbidden States

Either of the circuits described so far depicts an SR flip-flop, regardless of whether it is NAND-based or NOR-based. Its input and output states are summarized in Figure 11-7. However, as this table suggests, there are some input states that create problems.

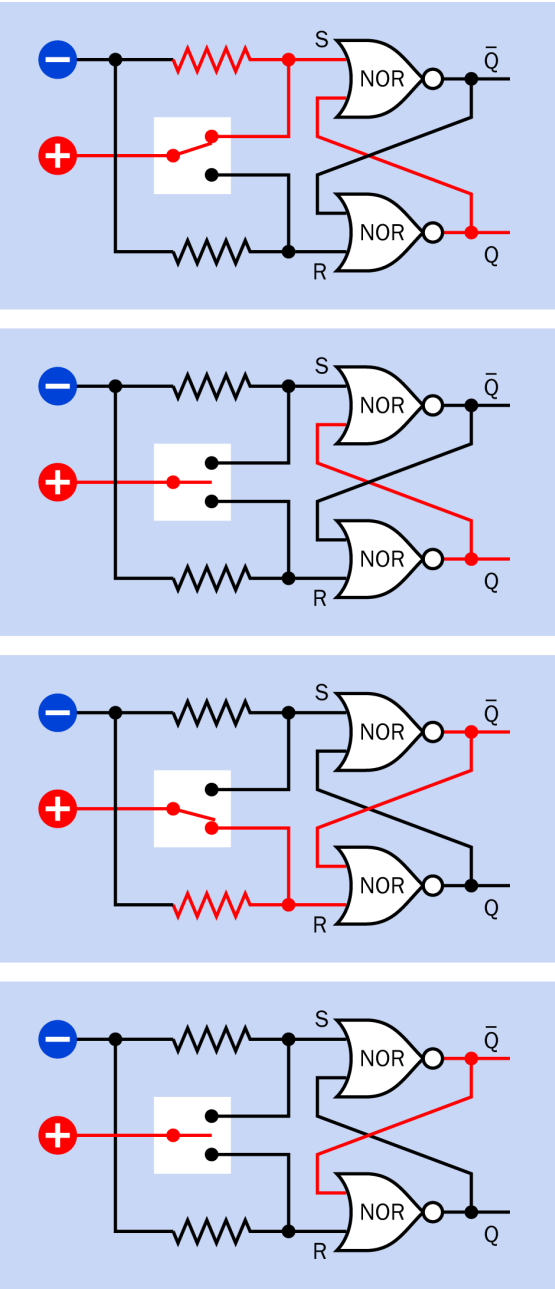


Figure 11-6. Four snapshots of a NOR-based SR flip-flop, showing the consequence of changing switch positions, comparable with the NAND-based flip-flop.

Flip-Flop Inputs		Flip-Flop Outputs			
		NAND-based		NOR-based	
S	R	Q	Q̄	Q	Q̄
●	●	● Problem ●		Same as Previous	
●	●	●	●	●	●
●	●	●	●	●	●
●	●	●	●	●	●
●	●	Same as Previous		● Problem ●	

Figure 11-7. A table of input states and the consequent output states for NAND-based and NOR-based SR flip-flops.

In either the NAND-based flip-flop or the NOR-based flip-flop, the output when the switch is in its unconnected center position will remain the same as when the switch was in its previous position. This is the usefulness of the flip-flop: it remembers the previous state. These situations are identified as “Same as Previous” in the table.

The pullup resistors (in a NAND-based flip-flop) and pulldown resistors (in a NOR-based flip-flop) are intended to guarantee that both inputs will be high (NAND) or both inputs will be low (NOR) even when the switch makes no connection. Therefore, it should be impossible for both inputs to be low (NAND) or high (NOR).

But what happens if the circuit is powered up with the switch in the unconnected position? One input of each gate is controlled by the output of the other gate. But what will those outputs be?

In the NAND-based version, the outputs from the NANDs will be low while the chips are powering up. As soon as the NAND chips are functioning, each of them will sense that it has one input high and one input low, so it will change its output to high.

But now that each chip has a high output, it will feed back to the secondary input of the other chip. Now both chips have both inputs high. This will cause them both to change their outputs to low—but this will feed back again, flipping the outputs back to high again. In fact, if the gates

are absolutely identical, the circuit will oscillate very rapidly. This is sometimes known as *ringing*.

In real life, the gates will not be absolutely identical, and eventually one of them will respond fractionally ahead of the other, tipping the circuit into the state shown either in the second panel or the fourth panel in Figure 11-4. But which chip will win? There is no way of knowing. This is known as a *race condition*, and the winner is unpredictable.

A similar but opposite situation occurs in the NOR-based flip-flop if it is powered up with the switch in the disconnected position, and the S and R outputs are both low, because of the pull-down resistors. Here again it will be a race condition.

We can address the problem by making a rule that the switch must always be in one position or the other when the flip-flop is powered up. But what if there is a faulty switch? Or what if a power interruption occurs while the switch is changing position?

Another problem occurs if the switch makes one contact a fraction before it breaks the other contact. This would result in both S and R inputs being low, in a NAND flip-flop. The same state could occur if a separate logic circuit is driving the S and R inputs, and an error causes it to make S and R both low. This is shown in Figure 11-8. Because the output from a NAND gate is always high if at least one of its inputs is low, both gates now have a high output, and the circuit is stable.

The problem is, the states of the outputs from a flip-flop should always be opposite to each other. If both of them are high, this can create logic problems in the rest of the circuit attached to the flip-flop.

- In a NAND-based SR flip-flop, if S and R are both low, this is known as a *forbidden state* or a *restricted combination*.

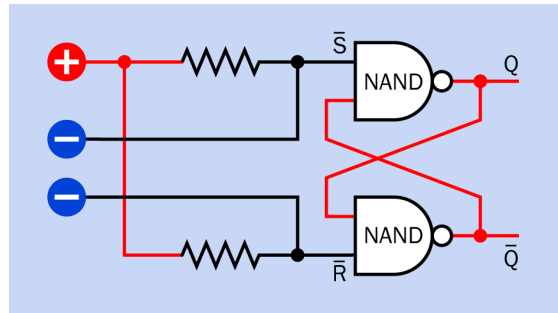


Figure 11-8. What happens when both S and R inputs to a NAND flip-flop are low as a result of an error in a separate control circuit.

A similar problem afflicts a NOR-based SR flip-flop, except that the forbidden state will occur when the S and R inputs are both high.

- In a NOR-based SR flip-flop, if S and R are both high, this is a *forbidden state* or a *restricted combination*.

The SR flip-flop is useful as a switch debouncer, but for computing applications, it is vulnerable to errors.

The JK Flip-Flop

Because the name of the JK flip-flop shares the initials of Jack Kilby, who won a Nobel prize for his fabrication of the world's first integrated circuit, some people speculate that this type of flip-flop was named after him. The attribution seems implausible, and may have gained currency simply because a flip-flop was the first device that Kilby happened to build when he was developing an integrated circuit.

Regardless of how it came to be named, the JK design is shown in Figure 11-9. This is commonly referred to as a JK *latch*. The electromechanical switches that were shown driving the SR flip-flop, along with pullup or pulldown resistors, are no longer included, because the inputs at positions J and K are assumed to come from other devices that have properly defined high and low states. Their behavior may be unpredictable, but neither of them will ever have a floating state.

This is a *gated* circuit, meaning that an additional input stage blocks direct access to the output stage, and it is also a *synchronous* circuit, as it uses a train of pulses at a clock input. Two three-input NAND gates are placed in front of a NAND-based SR flip-flop, and they address the problem of simultaneous identical inputs by using crossover feedback from the second stage to the first stage (via the conductors at top and bottom of the schematic).

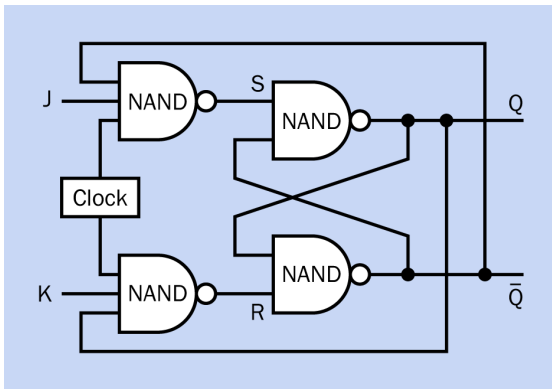


Figure 11-9. The basic circuit for a clocked JK flip-flop, using two additional NAND gates prepended to an SR flip-flop.

Versions of a JK flip-flop are possible using NOR gates, but are less common. Only the NAND-based version will be considered here.

In the case of a three-input NAND gate:

- All three inputs high: Output low.
- Other input combinations: Output high.

Because of the additional pair of NANDs, the circuit now recognizes a high input as logically active, instead of the low-active input in the previous SR flip-flop using NAND gates. Consequently, two simultaneous high inputs might be expected to create the type of forbidden state that was caused by two simultaneous low inputs previously. However, in [Figure 11-10](#), the top and bottom panels show that simultaneous high inputs at J and K will support two possible valid outputs, where the state at Q is always opposite to the

state at NOT-Q. In fact, when both inputs are high, a positive pulse on the clock input will *toggle* the outputs (i.e., they will switch places). In fact, the toggling will continue so long as the clock input is high. Consequently, this type of flip-flop is intended for use with short clock pulses.

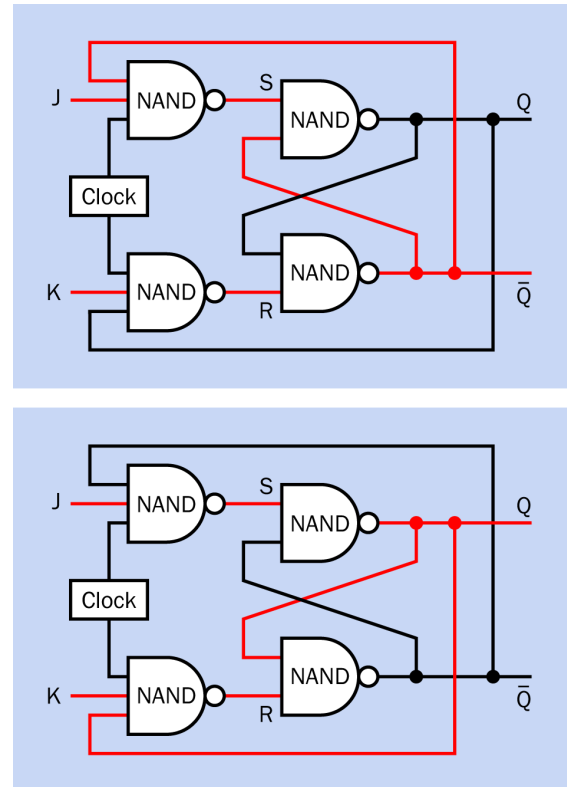


Figure 11-10. When the J input and the K input are both high, this no longer causes a forbidden state. The combination will toggle the outputs of the flip-flop between the two states shown here.

Master-Slave Flip-Flop

A more stable form is shown in [Figure 11-11](#) where yet another stage has been added, this one being a “master” to the first. In fact, this configuration is known as a *master-slave* flip-flop, the slave stage being driven by the master stage but remaining inactive until a low clock input at the master stage passes through an *inverter* to become a high clock input at the slave stage. The master and slave stages thus take turns, one be-

ing activated by a high clock pulse while the other is activated by the low part of the pulse cycle. The output from the slave stage cannot feed back to the master stage while the clock pulse is still high, and thus the timing issue in the single-stage JK latch is eliminated. Because the master-slave version of the JK configuration is not *transparent*, it is correctly known as a flip-flop rather than a latch.

In addition, Preset and Clear inputs may be added to override the clock to Set or Reset the outputs. These inputs are active-low.

Figure 11-12 summarizes the behavior of a JK master-slave flip-flop that is triggered by the falling edge of each clock pulse (shown as a downward-pointing arrow in the Clock column of the table). Note that the output will be delayed while the slave stage waits for the second part of each clock cycle.

The letter X in the table indicates that the state in that cell is irrelevant.

When J and K are both low, the states of Q and NOT-Q will remain the same as in the previous cycle, and this is still referred to as a *Hold* state. When J and K are both high, the outputs toggle, meaning that their new states will be opposite to the previous states.

D-Type Flip-Flops

AD-type flip-flop places an inverter between two inputs to guarantee that they will always be in opposite states, and uses a clock signal to copy their states to a pair of logic gates.

When an inverter is added between the inputs in this way, either an SR flip-flop or a JK flip-flop can become a D-type flip-flop. Figure 11-13 shows the simplest possible D-type circuit, added to a basic SR flip-flop. Only one data input is now required (customarily labeled D), because it drives the other through the inverter.

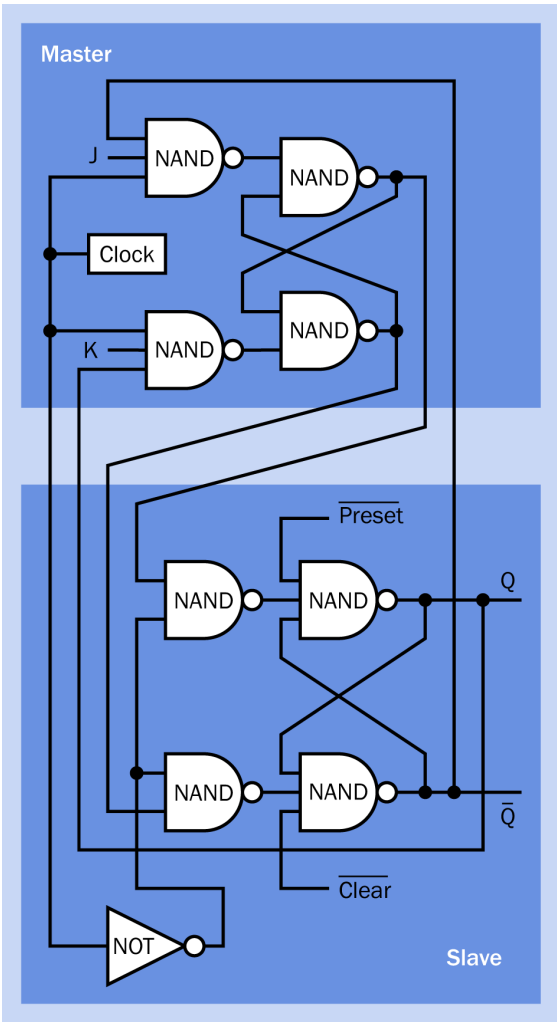


Figure 11-11. A master-slave circuit that drives one flip-flop with another.

Flip-Flop Inputs					Outputs	
Clock	Preset	Clear	J	K	Q	\overline{Q}
X	●	●	X	X	●	●
X	●	●	X	X	●	●
↓	●	●	●	●	Same as Prev	
↓	●	●	●	●	●	●
↓	●	●	●	●	●	●
↓	●	●	●	●	Toggle	

Figure 11-12. A table showing inputs and outputs for a JK master-slave flip-flop.

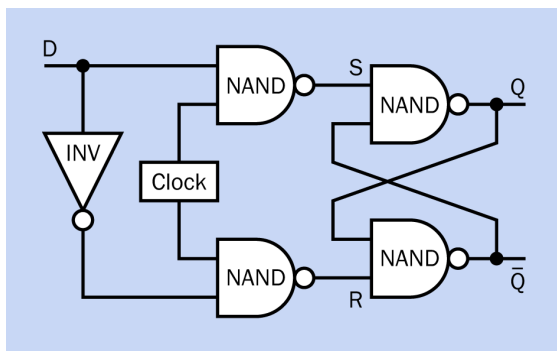


Figure 11-13. A simple D-type flip-flop. The inverter guarantees that the state of one input will always be opposite to the state of the other.

Figure 11-14 uses snapshots to show how the circuit responds to changing input and clock conditions. Initially, in the top panel, the data input is high, the clock input is high, and the Q output is high. In the second panel, the clock goes low, causing the output from the upper NAND gate in the input stage to change from low to high. But the upper NAND gate in the output stage still has one low input, so its state remains unchanged. In fact, the S and R inputs of the output NANDs are now both high, which creates the hold condition.

In the third panel, the D input changes from high to low, but this has no effect so long as the clock is low. The D input can fluctuate repeatedly, and nothing will happen until the clock goes high, as shown in the fourth panel. Now the clock copies the new D input state through to the output.

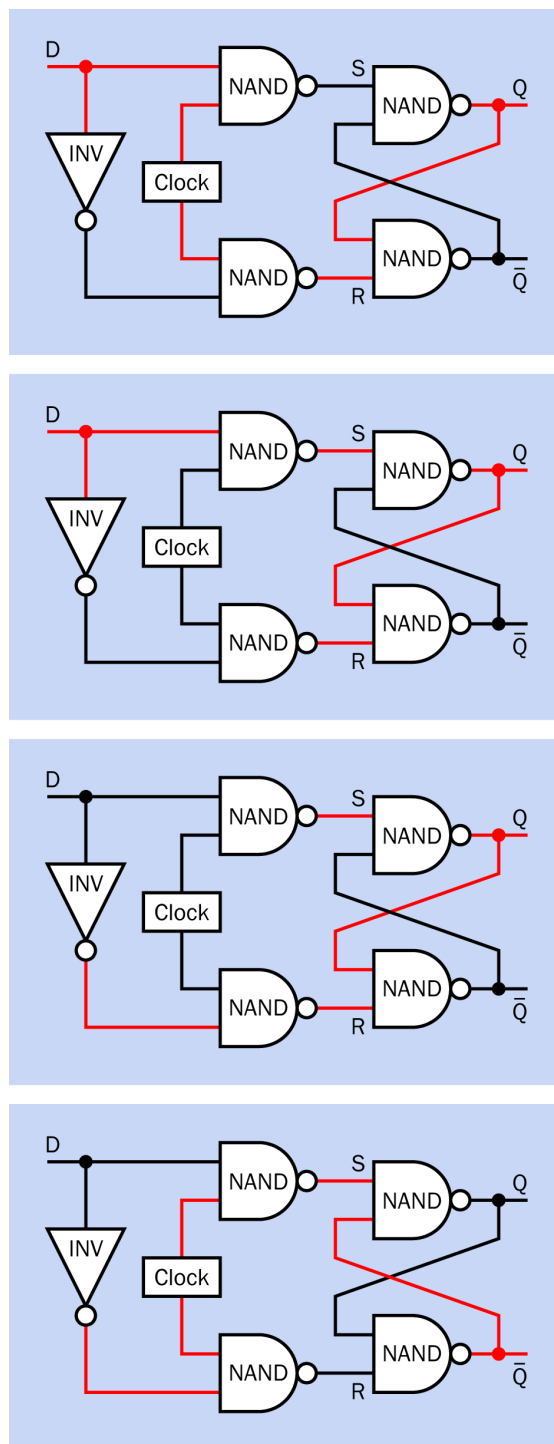


Figure 11-14. Four snapshots showing the behavior of a D-type flip-flop.

Summary

- An SR flip-flop can be used for switch debouncing, but in other applications it can enter an unacceptable race condition if its inputs and power supply are not carefully controlled.
- A JK flip-flop is gated, meaning that an SR circuit is preceded with an input stage and a clock input. This eliminates the race condition, adds the ability to toggle the outputs, but requires a very brief clock input. The circuit is edge-triggered.
- A master-slave flip-flop consists of two flip-flops, one driving the other. They can be JK type or SR type. The first flip-flop is activated by a positive clock state, while the second is activated by the subsequent negative clock state. Timing issues are resolved.
- A D-type flip-flop is gated with an inverter between the inputs, so that they cannot be simultaneously high or low. Consequently, only one input, labeled D, is needed. A high state on the D input causes a Set condition, while a low input causes a Reset condition, but only when the clock copies the status of the inputs through to the outputs. The status of the outputs remains stable (the flip-flop enters a hold condition) after the clock goes low.
- The JK circuit used to be widely used, because of its versatility. The D-type circuit now predominates.
- A T-type (toggling) flip-flop exists but is uncommon, and is not included in this encyclopedia.
- Many flip-flop circuits exist in addition to the ones that have been illustrated here. Only the most commonly cited circuits have been included.

A chip containing two positive-edge triggered D-type flip-flops is shown in [Figure 11-15](#). Each flip-flop in this component has its own data, set, and reset input and complementary outputs.

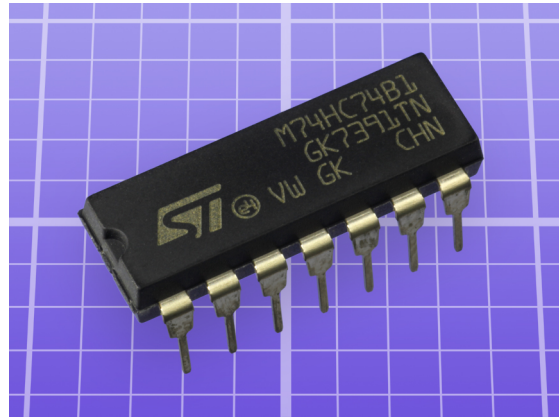


Figure 11-15. This chip contains two positive-edge triggered D-type flip-flops.

Variants

A selection of schematic symbols representing flip-flops is shown in [Figure 11-16](#). Letters S, R, J, K, or D define the type of flip-flop. Q and NOT-Q are the outputs. CLK is the clock input but may alternatively be identified with letter E, meaning *Enable*. SRCK or SCLK may also identify it, the abbreviations being intended to mean “serial clock.”

A triangle preceding CLK indicates that the flip-flop is positive-edge triggered. A circle, properly termed a *bubble*, preceding the triangle, indicates that the flip-flop is negative-edge triggered. In other locations, the bubble indicates that the input (or output) is inverted; it means the same thing as a bar printed above the text abbreviation, and indicates active-low logic. Synchronous inputs are shown on the left side of the flip-flop with the CLK input, while asynchronous inputs (if any) are shown above and below the flip-flop rectangle.

Using these guidelines, the examples in [Figure 11-16](#) can be decoded:

1. An unlocked SR flip-flop with active-low inputs (probably NAND-based).
2. An unlocked SR flip-flop with active-high inputs (probably NOR-based).

3. An SR flip-flop with active-high inputs, pulse-triggered by an active-high clock input.
4. A JK flip-flop with active-high inputs, edge-triggered by a rising-edge clock input. The bubble on the lower Q output means the same thing as a letter Q with a bar printed above it.
5. A D-type flip-flop pulse-triggered by an active-low clock input.
6. A D-type flip-flop edge-triggered by a falling-edge clock input.
7. A JK flip-flop with active-high inputs, edge-triggered by a rising-edge clock input, with asynchronous active-low Preset and Clear inputs.

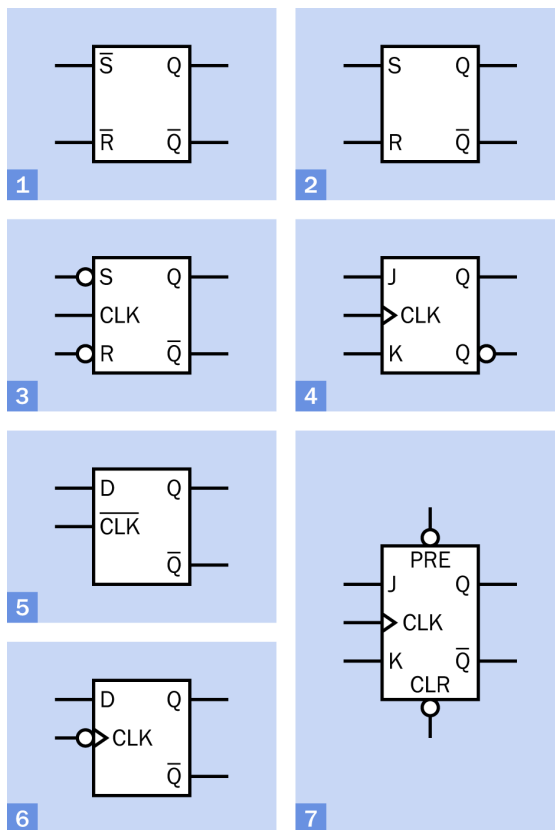


Figure 11-16. The schematic symbol for a flip-flop is an annotated rectangle. See text for an explanation of the letters and marks.

Packaging

Only about 10% of the flip-flops listed by a typical parts warehouse are through-hole chips. The rest are now surface-mount. Still, even if the search is narrowed further to through-hole packages in the 74xx and 4000 series, at least 100 types still exist. They provide opportunities in education and prototyping work, even though they are less often used as standalone components.

A package often contains more than one flip-flop. Dual and quad arrays are common. The flip-flops may be independently clocked, or may share a single clock input; datasheets should be checked carefully for details. Octal flip-flops, such as the D-type 74x273, are intended for use as eight-bit registers.

Many of the older flip-flops are numbered in the 74xx series of logic chips. See [Chapter 10](#) for a detailed guide to this numbering system and the various logic families. D-type flip-flops include 74x74, 74x75, 74x174, and 74x175, where an acronym for the logic family is substituted for the x. Old-style CMOS flip-flops include the 4042B D-type latch, the 4043B quad NOR SR flip-flop, and the 4044B quad NAND SR flip-flop. The last two are synchronous, and both allow two Set inputs, labeled S1 and S2 in the datasheet.

Examples of JK flip-flops include the 74x73, 74x76, and 74x109.

Values

As is the case with other logic chips, most flip-flops in the through-hole 74xx series are intended for 5VDC power supply while the older 4000 series may tolerate up to 18VDC. Surface-mount versions may use voltages as low as 2VDC.

See [“Variants” on page 93](#) for a discussion of acceptable high and low logic input voltages. On the output side, the 4000 series chips are able to source less than 1mA at 5VDC, but the 74HCxx series can manage around 20mA.

If a flip-flop is used for high-speed operation, the following values must be considered:

- t_s Setup time: The minimum time in nanoseconds for an input to be constant before the next clock pulse can process it.
- t_H Hold time: The minimum time in nanoseconds for an input to persist after the active edge of a clock pulse that has processed it. The interaction between a clock pulse and an input state takes a brief but measurable amount of time; errors may occur if the clock is given less than that amount of time to do its job.
- t_{CO} Clock-to-output: The elapsed time after an active clock edge, before the output changes. This is a function of the internal workings of the chip, and may be broken down into low-to-high and high-to-low output transitions, as follows.
- T_{PLH} Propagation to Low-to-High: The elapsed time after an active clock edge, before a low-to-high swing occurs at an output. This may not be identical to T_{PHL} .
- T_{PHL} Propagation to High-to-Low: The elapsed time after an active clock edge, before a high-to-low swing occurs at an output. This may not be identical to T_{PLH} .
- f_{MAX} Maximum clock frequency for reliable operation.
- $t_{W(H)}$ The minimum high clock pulse width in nanoseconds.
- $t_{W(L)}$ The minimum low clock pulse width in nanoseconds.

In a shift register or counter, where multiple flip-flops are cascaded but they share the same clock, the t_{CO} of one flip-flop must be shorter than the hold time of the next flip-flop, to allow the input of data to be complete before the window of opportunity is over.

How to Use It

The asynchronous SR flip-flop is of primary use in debouncing switches. Examples are the single MAX6816, dual MAX6817, and octal MAX6818.

D-type flip-flops are widely incorporated in frequency dividers, which are used to count pulses and display a binary output. If the NOT-Q output is wired back to the D input, the pulse stream to the clock input will have the following effect:

1. Suppose the initial D state is low and the initial state of the NOT-Q output is low.
2. The first high clock pulse propagates the low D state into the flip-flop.
3. The next low clock state forces the NOT-Q output high. This feeds back and creates a high D input.
4. The second clock pulse propagates the high D state into the flip-flop.
5. The next low clock state pulls the NOT-Q output low. This feeds back and creates a low D input.

The sequence then repeats. Only one high output is generated at NOT-Q (or at Q) for every two clock pulses; thus the circuit can become a divide-by-two **counter**. If the Q output is tapped to serve as the clock input for another flip-flop, that circuit now has a divide-by-four output. A series of many flip-flops can be chained together, so long as the propagation of signals along the chain is fast enough to occur before the next clock pulse. This is known as an *asynchronous counter*.

For more information on the use of counters, see [Chapter 13](#).

While flip-flops have tended to be integrated with other components in digital computing, they are still used as registers where 8 or 16 bits of serial data must be assembled at a time, prior to being disseminated as parallel data.

What Can Go Wrong

Ambiguous Documentation

For reasons which are unclear, instructional texts and tutorials can be erratic when describing flip-flops:

- A truth table may fail to clarify whether the circuit uses active-high or active-low logic.
- Truth tables from different sources are often inconsistent in their representation of current and future output states, and may even fail to include the clock status in a clocked flip-flop.
- Tutorials may include logic diagrams for some types of circuit, but not others.
- NOR gates may be used, without any mention that NAND gates can also be used (and may be more common or convenient).
- The active-low or active-high status of inputs in an SR flip-flop may not be shown.

Bearing this in mind, manufacturer datasheets should be consulted whenever possible as the primary source of information.

Faulty Triggering

In many cases, a flip-flop designed for edge triggering can give erroneous results if it is level-triggered, and vice versa. Rising-edge-triggered flip-flops must be distinguished from falling-edge-triggered flip-flops. As always, it is important for similar parts that have similar functions to be stored separately.

Metastability

The behavior of flip-flops has been described in this entry under ideal conditions, where they are

operating well within parameters established by the manufacturer. In reality, non-ideal scenarios may occur, especially where inputs such as data and clock, or clock and reset, are almost simultaneous. This may be difficult to avoid if a signal is received from an external source such as a sensor, with no way to control its arrival time. If the input occurs within the setup time or the hold time of a clock pulse, the flip-flop may be unable to determine whether the input precedes or follows the clock.

This may lead to *metastability*, meaning an unpredictable output and/or oscillations that take several clock cycles to settle into a stable state. If the output from a flip-flop may be used by two separate components with slightly different response times, one may interpret the oscillating output as a high state while the other interprets it as low. In a computing circuit, metastability can lead to calculation errors or a system crash. To avoid these issues, limits in datasheets should be observed. Attention should be paid to the manufacturer specifications for minimum setup time and hold time, so that the circuit has sufficient opportunity to recognize a signal and respond.

One solution to metastability is to connect multiple flip-flops in series, all sharing a common clock signal. This will tend to filter out irregularities, at the expense of requiring additional clock cycles if the flip-flops are not transparent.

Metastable-hardened flip-flops minimize metastability but cannot eliminate it completely.

Other Issues

Problems that tend to affect digital chips generally are listed in the section of the entry on logic gates (see [“What Can Go Wrong” on page 105](#)).

shift register

The term **shift register** is rarely hyphenated. In this encyclopedia, no hyphen is used.

A shift register can function as a *queue*, but this term is more usually applied to software. When the output from the last stage of a shift register is connected back to its input, it can function as a *ring counter*, but that application is described in the **counter** entry of this encyclopedia.

Component catalogs sometimes list shift registers as binary ripple **counters**, instead of giving them their own section. In this encyclopedia, a binary counter is considered to have binary-weighted outputs (with values 1, 2, 4, 8... in decimal notation) and is described in the **counter** entry. A shift register has outputs that are not necessarily binary-weighted.

OTHER RELATED COMPONENTS

- **flip-flop** (see [Chapter 11](#))
- **counter** (see [Chapter 13](#))
- **multiplexer** (see [Chapter 16](#))

What It Does

A *register* is a component (or a small section of computer memory) that stores information. The smallest unit of information is one *bit* (i.e., one *binary digit*) with a value 1 or 0 that can be represented by a high or low logic state. A **shift register** most commonly is designed to store eight bits, although some store four.

Each bit is memorized by the status of a **flip-flop** inside the register. For a detailed description of flip-flops, see [Chapter 11](#). When a pulse from an external *clock* is received by the shift register, all of the bits in storage are moved along one step, from each flip-flop to the next. The high or low status of an input pin at that moment is *clocked in* to the first flip-flop, while the bit in the last flip-flop is overwritten by the bit preceding it. A diagram representing the function of a basic four-bit shift register is shown in [Figure 12-1](#).

Note that the status of the input pin is ignored until the moment when a clock pulse copies it into the first flip-flop. In the figure, when the input pin has a brief high state that ends immediately before clock pulse three, the high state is ignored.

A shift-register chip is shown in [Figure 12-2](#).

Because the functionality of a shift register is now often incorporated in much larger logic chips, it is less widely used as a stand-alone component than it used to be. It is still useful for purposes of serial-parallel or parallel-serial conversion, and for small tasks such as scanning a matrix-encoded keyboard or keypad. It also has educational applications and can be used in conjunction with a microcontroller.

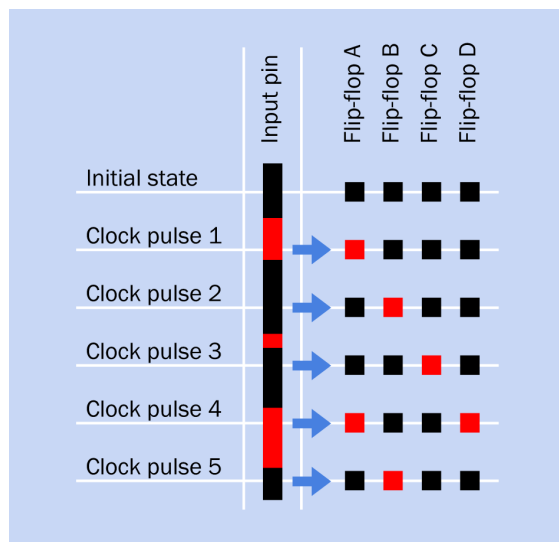


Figure 12-1. The function of a four-bit shift register in which each flip-flop may be set to a high or a low state, represented here with red or black squares. After a high bit is clocked into the chip, it is moved one space along by each subsequent clock pulse.

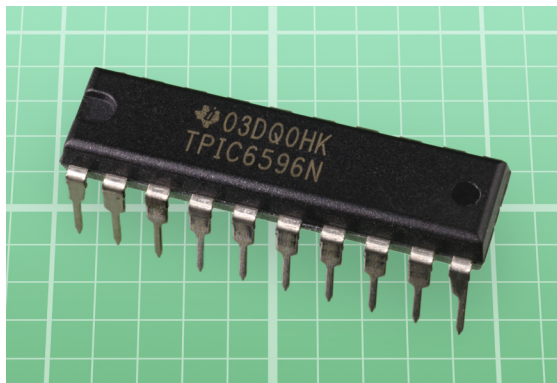


Figure 12-2. This 8-bit shift-register chip is unusual in that it uses “power logic,” in which open-drain outputs enable it to drive relatively high-current devices. It can sink up to 250mA at each of its output pins, at up to 45VDC.

Schematic Representation

No specific symbol exists for a shift register. It is represented in a schematic by a simple rectangle, often (but not always) with control inputs on the left, data inputs arrayed along the upper edge, and data outputs along the lower edge. An example is shown in [Figure 12-3](#), along with a dia-

gram showing the physical chip and its pinouts. The meaning of the abbreviations identifying the inputs, outputs, and control functions will be described in [“How It Works” on page 122](#).

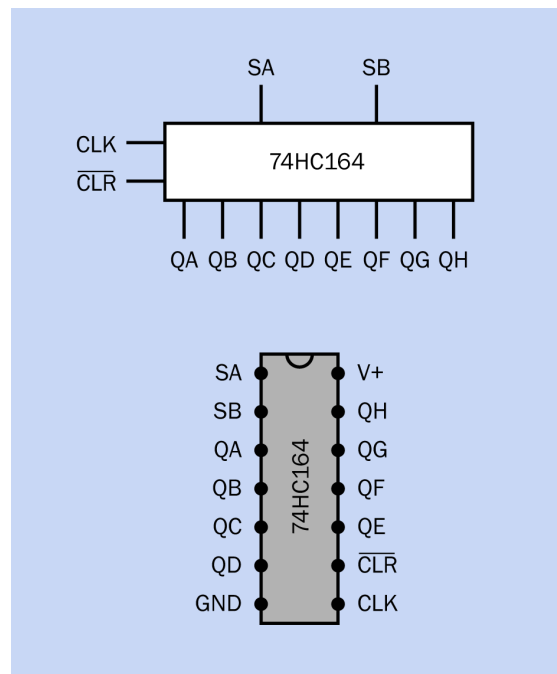


Figure 12-3. Typical schematic representation of a shift register, compared with the pinouts of the actual component.

The schematic symbol representing a shift register may appear superficially similar to the physical form of the chip which contains it, but the physical layout of the pins is unlikely to be the same.

How It Works

A shift register generally consists of a chain of D-type flip-flops. See the entry describing **flip-flops** in [Chapter 11](#) for a detailed explanation of this component.

The simplest shift register functions as a *serial-in, serial-out* device, abbreviated with the acronym *SISO*. Because the first bit that enters it will be the first to leave at the opposite end, it can also be

described as a *first-in, first-out* data storage device, using the acronym *FIFO*.

The basic connections between flip-flops in a four-bit SISO shift-register are shown in [Figure 12-4](#). The D input in each section refers to the fact that it is a D-type flip-flop. The primary output from each flip-flop is identified with letter Q.

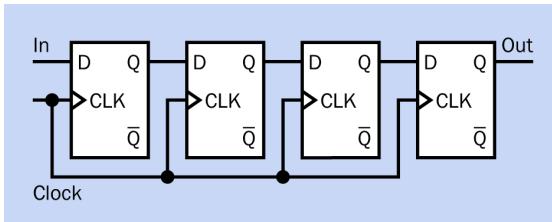


Figure 12-4. The simplest shift register is a serial-in, serial-out (SISO) device. This example contains four D-type flip-flops.

Each clock input is labeled CLK. When the output of each flip-flop is coupled to the input of the next flip-flop, and both share the same clock signal, the clock signal will cause the state of the third flip-flop to be sent to the fourth, the output of the second to be copied to the third, the output of the first to the second, and the input state will be copied to the first.

Abbreviations and Acronyms

The shift register will usually have an additional input that forces an immediate “clear” of all the registers, regardless of the clock state at that moment. This input is usually labeled CLR and will have a bar printed above it if it is active-low (which is the usual convention). If there is a pin labeled MR (meaning “Master Reset”), it will have the same function as CLR.

Because its effect is independent of the clock state, the clear signal is described as an *asynchronous* input.

While the abbreviation CLK is frequently used to identify the clock input, SCLK is also used (meaning “serial clock”), and occasionally the abbreviation CP may be found, meaning “clock pulse”

input. If the shift register contains two stages, one to clock data in and the second to clock data out, they may be separately clocked, in which case they will be identified with different abbreviations. These are not standardized, but should be explained in the manufacturer’s datasheet. No matter which abbreviation is used for a clock input, it will have a bar printed above it if the input is active-low.

Shift registers are generally *edge-triggered*, meaning that the rising or falling edge of a clock pulse triggers the bit-shifting operation. If the component responds to a clock transition from low to high, it is *rising-edge triggered*. If it responds to a transition from high to low, it is *falling-edge triggered*, and this may be indicated in the schematic by a small circle, properly known as a *bubble*, preceding the triangle which indicates that this is an edge-triggered device.

Most shift registers are positive-edge triggered.

Parallel Outputs and Inputs

In many shift registers, data may be read out in parallel (from all flip-flops simultaneously), using pins provided for this purpose. In this mode, the shift register can function as a *serial-parallel converter* (*serial in, parallel out*, represented by the acronym *SIPO*). A simplified schematic of the internal connections is shown in [Figure 12-5](#).

Where parallel outputs are provided, they are often identified as QA, QB, QC, and so on (moving from left to right) but may alternatively be described as Q1, Q2, Q3, Q4, and so on.

In a schematic, the input pin is conventionally shown as being at the left end of the component. Often two inputs are provided, connected internally as inputs to a NAND gate. The inputs are likely to be labeled A and B, but may alternatively be named SA and SB, indicating that they are serial inputs. S1 and S2 are alternative classifications. If parallel inputs exist, they may be identified as PA, PB, PC, and so on.

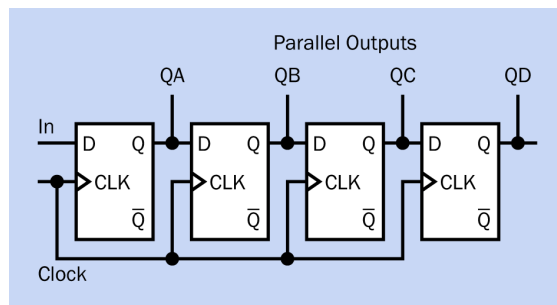


Figure 12-5. Many shift registers have pins connected to points along the chain of flip-flops. These connections enable data to be read from the shift register in parallel.

If serial data is supplied asynchronously, as in the illustration in [Figure 12-1](#), it will be ignored until the shift register is triggered by the next clock pulse. The input state at that moment will then be copied into the first flip-flop, while the data that is already being stored in the shift register will be moved along the chain. In datasheets, this is customarily represented by a diagram such as the one in [Figure 12-6](#). This diagram assumes that the shift-register is rising-edge triggered. Note that a brief fluctuation in the input which does not coincide with a clock-trigger event will be ignored.

Variants

Serial In, Serial Out

A basic SISO shift register allows only for serial input (at one end of the chain of flip-flops) and serial output (at the other end of the chain). No pins are available for parallel output of data.

This type of component usually permits 64-bit storage, where parallel output is simply impractical, as too many pins would be required. An example is the 4031B chip. This includes provision for recirculation of bits, so that it will also function as a ring counter (see [Chapter 13](#) for a discussion of this function). As is always the case with logic chips, the part number will be preceded by letter(s) identifying the manufacturer, and a suffix will distinguish variants of the chip.

Another type of SISO shift register is programmable. It will store any number of bits from 1 through 64, determined by a binary number applied in the form of high/low states to five pins reserved for this purpose. An example is the 4557B.

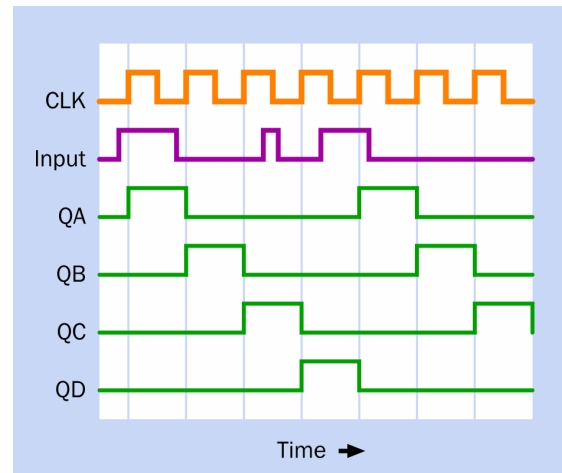


Figure 12-6. In a rising-edge-triggered shift register, the high or low state of an asynchronous input (purple line) is copied into the first flip-flop of a shift register by each clock pulse (orange line). Brief fluctuations that do not coincide with a rising clock pulse are ignored. Existing data in the register is shifted from one flip-flop to the next.

Serial In, Parallel Out

The majority of serial-input shift registers allow parallel output from points along the chain in addition to serial output at the end of the chain. These chips almost all are 8-bit registers. Typically two inputs are provided, one of which can be used to receive bits that recirculate from the end of the chain, back to the beginning. Widely used examples are the 4094B and the 74x164, where an acronym identifying the logic family will be substituted for the x.

Parallel In, Serial Out

A minority of shift registers are able to function as *parallel-serial converters* (*parallel in, serial out*, represented by the acronym *PISO*). Typically this type of chip allows *jam-type* parallel data input, meaning that the data is forced into the flip-flops

via a separate pin for each of them. Parallel input is enabled by the status of a serial/parallel control pin. When the control pin reverts to its opposite status, each clock pulse will now shift data along the chain of flip-flops, allowing it to be read from the final output one bit at a time. Thus, data can be entered into the chip in parallel and read out of it serially. Examples are the 4014B and 4021B. Both are 8-bit.

Parallel In, Parallel Out

Shift registers that permit parallel output in addition to parallel input are almost all of the *universal* type, described in the next section.

Universal

A *universal* shift register is capable of all four modes of operation: SISO, SIPO, PISO, and PIPO. The four modes of the component are selected by the high or low status of two *mode select* pins. In addition, this component may have the ability to shift the register states either left or right. A *bidirectional* shift register has this same capability, and may also have PISO and PIPO capability, depending on the chip. Examples are the 74x195 and 74x299, where an acronym identifying the logic family will be substituted for the x in the number.

Universal shift registers are almost all 4-bit or 8-bit. They often have relatively complicated features, such as access to internal JK flip-flops, or pins that are multiplexed to provide different functionality depending whether an enable pin is held high or low. Datasheets must be checked carefully to ensure correct use.

Dedicated shift registers of SIPO or PISO type will be easier to use.

Values

As is the case with other logic chips, most flip-flops in the through-hole 74xx series are intended for 5VDC power supply while the older 4000 series may tolerate up to 18VDC. Surface-mount versions may use voltages as low as 2VDC.

See the section on logic gates in [Chapter 10](#) for a discussion of acceptable high and low logic-input states. On the output side, the 4000 series chips are able to source less than 1mA at 5VDC, but the 74HCxx series can usually manage around 20mA.

If a shift register is used for high-speed operation, the following values must be considered (identical notation, and similar values, are found in specifications for flip-flops):

- t_S Setup time: The input state of a shift register must exist for a very brief period before the clock trigger that processes it. This period is known as the *setup time*. In the 4000 series of chips, recommended setup may be as long as 120ns. The value will be much lower in 74xx chips.
- t_H Hold time: The minimum time in nanoseconds for an input to persist after the active edge of a clock pulse that has processed it. In many shift registers, no hold time is necessary, as the chip has already been activated by the rising edge of the clock pulse.
- t_{CO} Clock-to-output: The elapsed time after a clock trigger, before the output changes. This is a function of the internal workings of the chip, and may be broken down into low-to-high and high-to-low output transitions, as follows.
- T_{PLH} Propagation to Low-to-High: The elapsed time after an active clock trigger, before a low-to-high swing occurs at an output. This may not be identical to T_{PHL} .
- T_{PHL} Propagation to High-to-Low: The elapsed time after an active clock trigger, before a high-to-low swing occurs at an output. This may not be identical to T_{PLH} .
- f_{MAX} Maximum clock frequency for reliable operation. In the older design of 4000 series chips, 3MHz may be recommended with a power supply of 5VDC. Higher frequencies are possible with a higher voltage power

supply. Frequencies as high as 20MHz are possible in the 5VDC 74HC00 series.

- $t_{W(H)}$ The minimum high clock pulse width in nanoseconds. In the older design of 4000 series chips, 180ns may be recommended with a power supply of 5VDC. Shorter pulses are possible with higher voltage power supply. Pulses as short as 20ns are possible in the 5VDC 74HC00 series.
- $t_{W(L)}$ The minimum low clock pulse width in nanoseconds. This is likely to be the same as $t_{W(H)}$.

Power Considerations

Shift registers conform with the usual power-supply requirements for logic families. These are described in detail in the **logic gate** entry in [Chapter 10](#). Likewise, the ability of a shift register to source or sink current is usually determined by its logic family. In a few cases, however, shift registers have an additional *open-drain* output stage for each register, capable of sinking currents as high as 250mA. The Texas Instruments TPIC6596 shown in [Figure 12-2](#) is an example. When an open-drain output is connected with a logic device whose input cannot be allowed to float indeterminately, a *pullup resistor* must be added.

Three-State Output

A chip may have a *three-state output* (also known as *tri-state output*, a term which was trademarked but is now used generically). This means it will be capable of changing its outputs so that instead of sourcing or sinking current in a logical high or low state, they can have a high impedance. The chip then becomes “invisible” to others which may be sharing the same output bus. The high-impedance state is usually applied to all outputs from the shift register simultaneously when enabled by a separate pin, often identified as *OE*, meaning *output-enable*. Examples of three-state shift registers are the 74x595 or the 4094B chip.

The high-impedance state can be thought of as being almost equivalent to switching the out-

puts of the shift register out of the circuit. Consequently, if other components sharing the bus are also in high-impedance output mode, the state of the bus will “float,” with uncertain results. To avoid this, a pullup resistor of 10K to 100K can be used on each bus-line.

Where the internal components of a shift register are shown in a datasheet, a three-state output is usually represented with a buffer symbol or inverter symbol that has an additional control input located on its upper edge, as shown in [Figure 12-7](#). This should not be confused with the similar placement of a positive power supply input to an amplifier or op-amp. (Schematics showing the interior elements of a logic chip almost never include power-supply connections.)

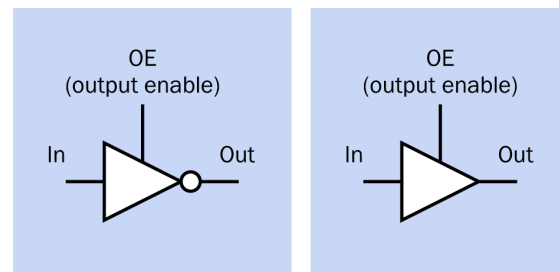


Figure 12-7. A shift register may be capable of a *three-state* output, where high impedance is an option additional to the usual high or low logic state. An output enable pin allows this. It is typically shown as an additional input to an inverter (left) or buffer (right) inside the shift-register chip.

How to Use It

The SISO application of a shift register can be used simply to delay the transmission of data by storing it and moving it from one flip-flop to the next before it is read out of the end of the chain.

The SIPO application of a shift register (serial in, parallel out) may be useful where a microcontroller has insufficient outputs to control multiple devices. Serial data can be sent from a single microcontroller output to the input of a shift register. The chip can then drive a separate device from each of its parallel output pins. If there are eight devices, the microcontroller can send a se-

quence of eight bits, each of which will control the on/off status of one device when the bits are read out of the shift register in parallel. If more devices are used, an additional shift-register can be daisy-chained to the output from the first.

Clock signals can be supplied from the microcontroller, along with a signal to the clear input of the shift register, if desired. Alternatively, the old bit states in the shift register can simply be “clocked out” and replaced with a new set of serial data. During the process of “bit banging,” the parallel outputs of the shift register can remain connected directly with the output devices if the clock speed from the microcontroller is fast enough, as devices such as relays will not respond to extremely brief pulses.

For generic shift registers that do not have open-drain outputs, a buffer will be needed to provide sufficient current for any device using more current than an LED.

If a shift register is configured for PISO mode (parallel in, serial out) it can be placed on the input side of a microcontroller, polling a variety of devices and feeding their states into the microcontroller serially.

Dual Inputs

Where a shift register has two serial inputs (as is often the case), they are almost always linked as inputs to an internal NAND gate. This allows the output from the end of the chain of flip-flops to be connected back to the beginning of the chain, if the shift register is to function as a ring counter. However, if this function is not used and a single input is required, the two inputs to the shift register can be tied together. In this configuration, the input becomes active-low. The two possible configurations are shown in Figure 12-8. It is important never to leave an input *floating*, or unconnected.

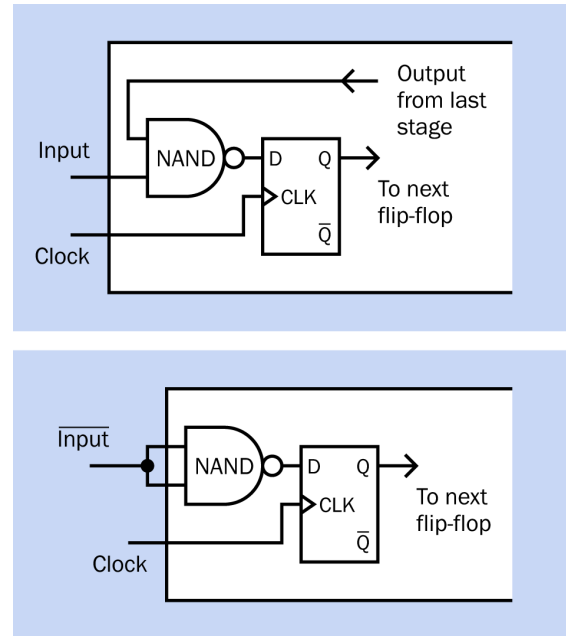


Figure 12-8. Two possible configurations where a shift register allows two inputs linked with an internal NAND gate.

Preloading the Shift Register

Where a shift register will be used to output a single recirculating bit (or in similar applications), the component must be preloaded with a high state in its first register. This may require a one-shot (monostable) timer which is activated only when the circuit is powered up.

Polling a Keyboard

Two shift registers can be used to scan the data lines in a matrix-encoded keyboard or keypad. This is often known as *polling* the keyboard. Provided the scan rate is sufficiently fast, the user experiences a seemingly immediate response to a key-press.

While the full schematic is too large and complex to be included here, many examples can be found online.

Arithmetical Operations

Shift registers were traditionally used to perform arithmetical operations on binary numbers. If the

number is represented by eight bits (i.e., one byte) with the most significant digit on the left, shifting the bits one space to the right will have the effect of dividing the byte value by 2. If the bits are shifted one space to the left, the byte value will be multiplied by 2 (provided an additional register is available to store the most significant bit after it has been shifted). This concept is illustrated in [Figure 12-9](#).

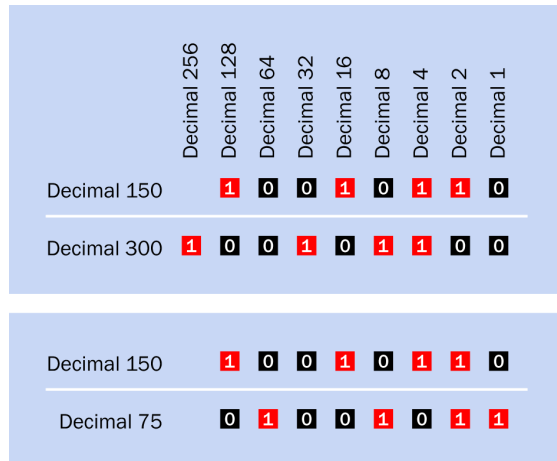


Figure 12-9. In the upper section of this diagram, a binary number represented by eight bits in a shift register is multiplied by 2 by shifting all the bits one space to the left. In the lower section, the same binary number is divided by 2 by shifting all the bits one space to the right. The binary values are shown in decimal notation to the left.

In the upper section of the figure, the binary number 10010110 (chosen arbitrarily) is represented in the eight flip-flops of a shift register. The decimal place value of each digit is indicated. Adding up the place values of the 1s in the number, the total is $128 + 16 + 4 + 2 = 150$. Below the white line, the digits have been shifted one place to the left, with the leftmost digit carried over to an additional location, while a zero is inserted in the rightmost location. Assuming that the additional place at the leftmost location has a place value of 256, the total is now $256 + 32 + 8 + 4 = 300$.

In the lower section of the figure, the same binary number has been shifted one space to the right,

with a 0 introduced in the leftmost location. The decimal value is now $64 + 8 + 2 + 1 = 75$.

While this application for shift registers was common during development of digital computing in the 1960s and 1970s, the shift register as a separate component became less common subsequently, as its functionality was acquired by modern CPU chips.

Buffering

A shift register may also be used as a *buffer* between two circuits where the clock speeds are different. Digits are clocked in by the first circuit, then clocked out by the second. Some shift registers allow two clock inputs and can be used for this purpose.

What Can Go Wrong

Problems that tend to affect digital chips generally are listed in the entry on logic gates (see [“What Can Go Wrong” on page 105](#)).

Confusing Classification

Because of the functional similarity to a binary ripple **counter**, a shift register is sometimes listed by component suppliers as a counter. In fact, a binary counter will almost always have outputs that have place values 1, 2, 4, 8.... and upward, while the outputs from a shift register will not have place values.

When searching for a shift register, it can be found by specifying a “counting sequence” of serial to parallel, serial to serial, parallel to serial, or parallel to parallel. If the “counting sequence” is simply up or down, the component is a counter, not a shift register.

Inadequate Setup Time

Each flip-flop in a shift register must have a stable input state before the next triggering event shifts the data. If this setup time is reduced below the minimum specified in the datasheet, results will be unpredictable.

Unconnected Input

Because many shift registers have a choice of two inputs to the same chain of internal flip-flops, it is easy to leave one of them unconnected by accident. A floating input will be vulnerable to any stray electromagnetic fields, and is almost certain to create random effects.

Output Enable Issues

The output enable pin on a shift register that has three-state logic outputs is usually active-low. Consequently, if the pin is left unconnected, the logic outputs may go into high-impedance

mode, or will fluctuate unpredictably. Where three-state outputs are not required, a safe course of action will be to avoid using three-state chips.

Floating Output Bus

If a pullup resistor is omitted from a bus that is shared by three-state chips, the results will be unpredictable. Even if the circuit design seems to guarantee that at least one chip will have a high or low output on the bus, a pullup resistor should still be included.

counter

13

The term **counter** is used here to mean a digital-logic chip. A counter could be built from discrete transistors, but this approach is obsolete. Counters may also be devised from parts such as multiple relays, or a solenoid advancing a ratchet wheel, but such electro-mechanical devices are not included in this encyclopedia.

In this encyclopedia, a counter by definition has binary-weighted outputs with values 1, 2, 4, 8, ... in decimal notation. The exception to this rule is a *ring counter*, which does not have binary-weighted outputs but is included here because its name identifies it as a counter. A **shift register** may be used as a ring counter, but is more versatile and has many other functions; hence it has a separate entry.

Gray code counters, in which successive outputs differ by only one binary digit, are not described in this encyclopedia.

OTHER RELATED COMPONENTS

- **flip-flop** (see [Chapter 11](#))
- **shift register** (see [Chapter 12](#))

What It Does

A counter can be used to count events, or can measure time in convenient intervals if it is connected with a component such as a *quartz crystal* that operates at a precise and reliable frequency. The counter receives input pulses (usually referred to as a *clock input*) and counts a predetermined number of them before restarting from the beginning. It will repeat in this fashion so long as power is connected, and the clock pulses continue, and a reset signal is not supplied.

Almost all counters create some form of output during the count. Most commonly, the output is a pattern of high and low states expressing the number of clock pulses in binary code. Where a counter counts to a very high number before recycling, some intermediate binary digits may be omitted.

While standalone counter chips are not as widely used now as in the early days of computing, they still find application in industrial processes, small devices, and education, and can be used to control incremental devices such as stepper motors. They can be used in conjunction with microcontrollers.

Schematic Representation

No specific logic symbol exists for a counter. It is most often shown in a schematic as a rectangle with clock input(s) and clear input(s) on the left and outputs on the right. An example appears in [Figure 13-1](#), above a representation of the physical chip and its pinouts. The meaning of the abbreviations identifying the inputs, outputs, and control functions will be found in [“How It Works” on page 132](#). Because the two MR inputs for this particular counter are ANDed inside the chip,

the AND symbol is included with the counter symbol.

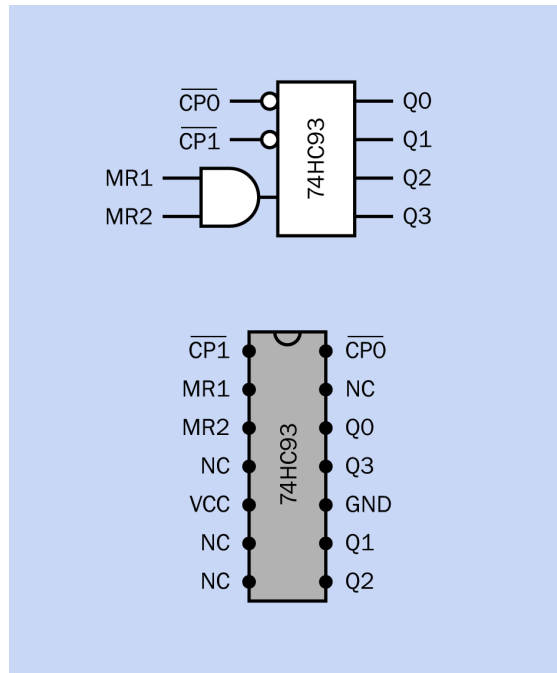


Figure 13-1. Typical schematic representation of a counter, compared with the pinouts of the actual component.

A counter chip is shown in [Figure 13-2](#).

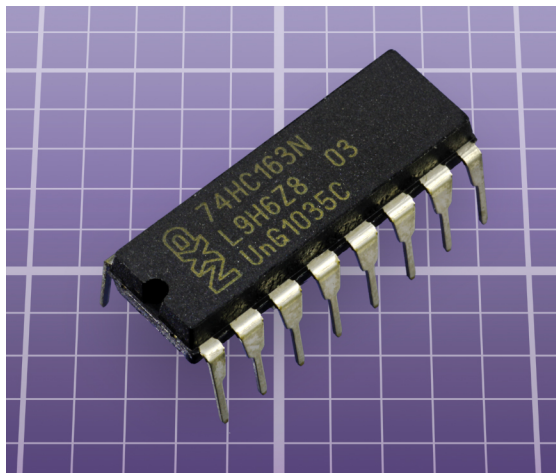


Figure 13-2. The 74HC163 shown in this photograph is a 4-bit synchronous counter capable of being preloaded with a starting value, and able to do a synchronous reset.

How It Works

A counter is built from a chain of **flip-flops**, with each one triggering the next. JK, T-type, or D-type flip-flops may be used. For a thorough description of a flip-flop, see [Chapter 11](#). In [Figure 13-3](#), a D-type flip-flop is shown, triggered by each rising clock pulse.

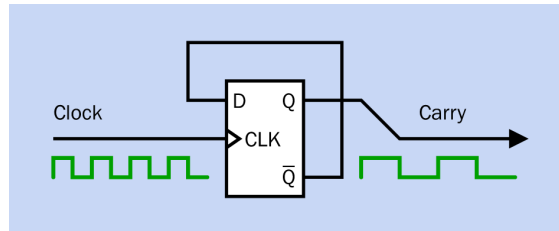


Figure 13-3. When the complementary (NOT-Q) output from a D-type flip-flop is connected back to its input, the Q output frequency is half of the clock input frequency. See text for additional details.

Initially the Q output of the flip-flop is low, so its NOT-Q output (identified by Q with a bar above it) is high. This feeds back to the D input, but has no effect until the rising edge of the next clock pulse copies the high D input to the Q output. The Q output is now latched high while the NOT-Q output is latched low and feeds back to the D input. The triggering event has passed, so the low D input does not have any immediate effect. The rising edge of the next clock pulse copies the low state of the D input to the output, and changes the NOT-Q output to high, causing the cycle to repeat. As a result, the output frequency of the flip-flop is one-half of the input frequency from the clock. If the output is carried to the next flip-flop to become its clock input, once again the frequency will be divided by two.

Modulus and Modulo

The *modulus* of a counter is the number it will count up to, before it repeats. This is sometimes written incorrectly as *modulo*.

In fact, *modulo* is the name of an arithmetical operation, often abbreviated as *MOD* (usually capitalized, even though it is not an acronym). This

operation consists of a division sum in which the *remainder* from the division is the result of the operation. Thus 100 modulo 5 gives a result of 0, because there is no remainder when 100 is divided by 5. But 100 modulo 7 gives a result of 2, because 2 is the remainder of the division operation.

To exacerbate the confusion, *MOD* is also used as an adjective referring to the modulus of a counter. Thus a MOD-4 counter has a modulus of 4, and a MOD-16 counter has a modulus of 16. As a general rule, when a counter is being described, *modulo* and *MOD* will mean the same thing as *modulus*. This may be confusing to people such as computer programmers who are already familiar with the correct usage of MOD as an arithmetical operator.

In a counter, to obtain a modulus that is not a power of two, logic gates inside the chip can intercept a particular value (such as 1010 binary, which is 10 decimal) and use this as a signal to restart the count at zero. External connections to the chip can achieve the same purpose.

Pin Identifiers

Abbreviations and acronyms are used in data-sheets to identify pin functions. These identifiers have not been standardized, and many variants exist.

CLK is the abbreviation most commonly used for the clock input, sometimes alternatively shown as CK or CP. If it is active-low, or if its falling edge will be active, a bar will be printed above it. Where a printed font does not permit an underscore to be placed in this way, CLK' (the abbreviation followed by an apostrophe) may be used instead. Often two or more stages of a counter can be separately clocked, and the input pins will be identified with abbreviations such as CLK1 and CLK2, or 1CLK and 2CLK, or CKA and CKB, or CP1 and CP2, or similar.

Where a clock input is edge-triggered, this is indicated by a small triangle. The triangle can be seen in [Figure 13-3](#).

CLR designates a pin which will clear the count and reset it back to zero. This signal is often active-low, indicated by a bar printed above the abbreviation.

In a schematic, a small circle, properly known as a *bubble*, may be placed at any input which is active-low. On a clock input, the circle indicates that it is falling-edge triggered. See [Figure 11-16](#) for various implementations of symbols with flip-flop schematics.

The CLR operation may be *synchronous* (the pin state will not be recognized until the next clock pulse) or *asynchronous* (in which case the pin state overrides the clock and resets the counter immediately). MR stands for "Master Reset," and has the same function as CLR.

Where two or more counters (or multiple stages within one counter) can be reset separately, more than one clear input will be shown, and may be identified with abbreviations such as CLR1 and CLR2, or MR1 and MR2.

Output pins will almost always be identified as Q0, Q1, Q2 ... or QA, QB, QC ... up to the maximum necessary to express the modulus. If two or more counters are included in one chip, outputs may be prefixed with an appropriate number. Thus 2Q3 would be the third output in the second counter. Multiple counters in one chip are identified with numbers starting from 1.

Where internal flip-flops are shown, they will have identifiers such as FF1 or FF2. Each flip-flop will have its own clear function, identified as C or CD (the latter meaning "clear data"), and may have inputs labeled D1, D2, D3... in a D-type flip-flop or J and K in a JK flip-flop. See [Chapter 11](#) for an explanation of flip-flop inputs and outputs.

The input to a counter is always imagined to begin from the left, and consequently the leftmost flip-flop shown in an internal schematic will express the least significant bit of the current value of the count, even though a binary number is written with the least significant bit in the rightmost place.

If a counter is capable of receiving parallel data as an input (explained below) it will have a pin labeled PE for parallel enable. It may also have a CE or CET pin, for count enable.

As is generally the case in logic chips, VCC or V+ are typically used to identify the positive power supply pin, while GND or V- will identify the negative-ground pin. NC means that a pin has no internal connection at all, and consequently requires no external connection.

Variants

All counter chips use binary code internally, and the number of *bits* (*binary digits*) in the counter's modulus will be the same as the number of internal flip-flops. A 4-bit counter (the usual minimum) will have a modulus of 2^4 which is 16. A 21-bit counter (the maximum typically available) will have a modulus of 2^{21} which is 2,097,152. For higher moduli, counters can be chained together, each sending a carry signal to the next. This is known as a *cascade*.

Multiple counters, with different moduli, may be chained in a single chip. For example, in a digital clock that displays hours and minutes using the 60Hz frequency of an American domestic power supply as its timebase, the initial counting stage will have a modulus of 60, to count individual seconds. The next counting stages will have moduli of 60, 10, and 6, so that they can count from 00 to 59 minutes. Additional stages in the chip will tally hours.

A counter with a *parallel input* can be preloaded with an initial value (in binary code) from which it may count up or down. A *parallel-enable* pin may put the counter into a mode where the number can be *jam loaded*, meaning that it is jammed into the counter regardless of the clock state. Other types of counter are loaded synchronously.

Ripple versus Synchronous

In a *ripple counter* each internal flip-flop triggers the clock input of the next, so that their states

change in a rapid but incremental sequence from the input to the output. This is also known as an *asynchronous* counter. Because the final state will not be valid until the clock pulse has rippled all the way through the counter (and through additional counters if they are cascaded together), a ripple counter will tend to suffer from a *propagation delay* of up to a microsecond. Ripple counters may also create output spikes or momentary transient count values that are invalid. Therefore, they are more suitable for applications such as driving a numeric display than for interfacing at high speed with other logic chips.

In a *synchronous counter*, all the flip-flops are clocked simultaneously. A synchronous counter is better suited to operation at high speed.

Of the counter chips available today, about half are synchronous and half are asynchronous.

Ring, Binary, and BCD

A counter that activates output pins one at a time sequentially is said to have a *decoded output*. It is often referred to as a *ring counter*. It has the same number of output pins as its modulus. An example is the 4017B chip.

A *binary counter* is more common and has an *encoded output*, meaning that it will express the running total of the count in binary code through *weighted outputs* that typically have (decimal) values of 1, 2, 4, 8, and so on. A modulus-8 counter (often referred to as an *octal counter*) will require three outputs which represent the binary numbers 000, 001, 010, 011, 100, 101, 110, and 111 (decimal 0 through 7) before going back to 000 and repeating.

A *modulus-16* counter, also known as a *hexadecimal counter* or a *divide-by-16 counter*, will have a binary output represented by four output pins, counting from 0000 through 1111 (decimal 0 through 15). Four-digit binary counters are very common, and their outputs are compatible with other components such as a **decoder**, which converts a binary-number input into a ring-counter-style output.

A **decade counter** is a modulus-10 binary counter. It is described as having a **binary-coded decimal** output (often expressed with the acronym **BCD**), using four weighted output pins which represent the numbers 0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, and 1001 (decimal 0 through 9) before repeating. Because this counter skips binary outputs from 1010 through 1111 (decimal 10 through 15), it is said to have a **shortened modulus**.

Figure 13-4 shows a schematic diagram of JK flip-flops in a decade ripple counter. The J and K inputs are all tied to the positive power supply, as this causes the clock input to toggle the output high and low. Note that because the primary input is always shown at the left end of the component, the least significant output bit (Q0) is in the leftmost position.

To intercept binary 1010 (decimal 10), an internal NAND gate is used. Its output goes low when its two inputs, from Q1 and Q3, go high. The output from the NAND immediately activates the CLR function on all the flip-flops, so that as soon as the decade counter reaches 1010 (decimal 10), it resets itself to 0.

In this particular chip, the preload for each flip-flop is tied to the positive power supply, so that it is always inactive. In some counters, the preload feature of each flip-flop is accessible via pins outside the chip. This creates the potential hazard of preloading the counter with one of the numbers that it normally skips (for instance, 11 decimal in a decade counter). This is referred to as an **invalid number** or **disallowed state**. (This use of the term “state” refers only to the binary number stored in the counter’s flip-flops. It has nothing to do with the high-state or low-state voltages used to represent binary 0 or 1.)

The counter’s datasheet should include a **state diagram** showing how the counter will deal with this situation. It may reset itself to a valid value after a maximum of two steps, but this can still cause confusion, depending on the application.

The state diagram for a 74HC192 counter is shown in Figure 13-5.

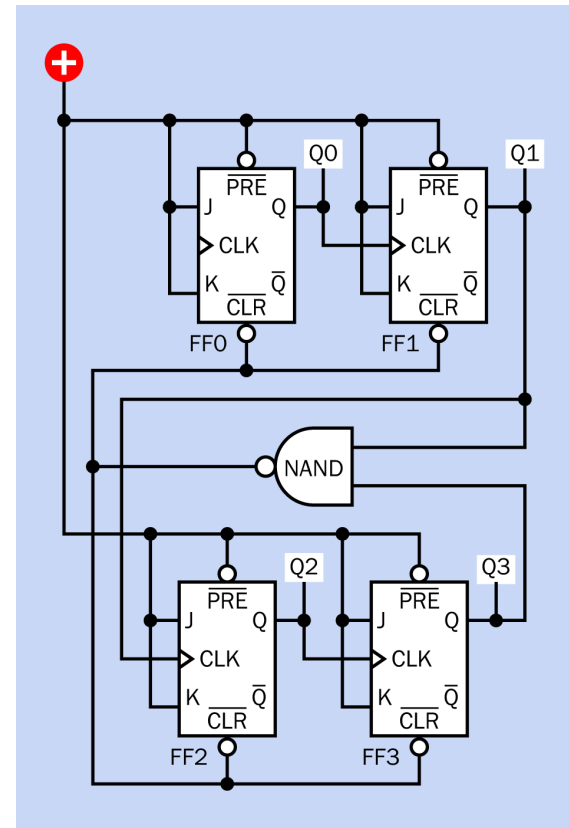


Figure 13-4. The internal logic of a synchronous decade counter that uses JK flip-flops.

Clock Sources

The clock input may be provided by a **timer** chip or by an **RC network**, which has the advantage of being able to run at a relatively low speed for purposes where this is desirable. It may alternatively be provided by a **quartz crystal** oscillating at a much higher frequency such as 1MHz. Successive counters may be necessary to reduce this value, depending on the application.

In some counters, the clock is built into the chip. More commonly, a resistor and capacitor may be used externally to establish a clock rate in conjunction with logic gates inside the chip. The datasheet for this type of component will include

a formula for calculating the clock frequency from the resistor and capacitor values. The 4060B chip is an example.

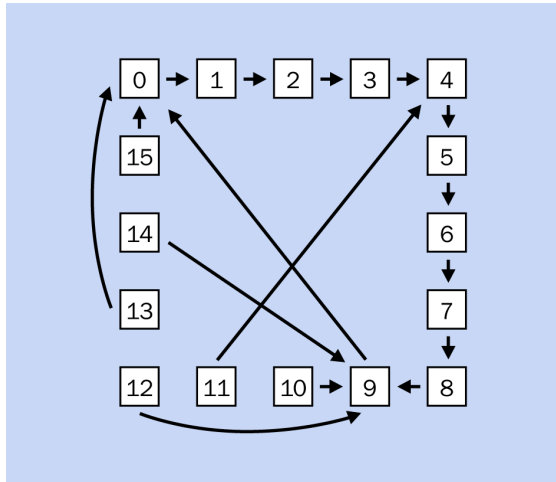


Figure 13-5. A state diagram shows the transitions that a counter will make from each number to the next (in decimal notation), including the transitions which enable it to exit from disallowed states. This example is for a 74HC192 chip.

Rising Edge and Falling Edge

A counter may be designed to be triggered either by the rising edge or the falling edge of the clock input, or by its high or low logic state. Generally speaking, ripple counters use the falling edge, so that the final output from one counter can become the clock input of the next. In other words, when the most significant digit of the first counter changes from a high to low logic state, this transition toggles the least significant bit of the second counter.

Synchronous counters generally use the rising edge of the clock input. If multiple synchronous counters are cascaded, they must all share the same clock signal, and will all change their flip-flop states simultaneously.

Multiple Stages

It is common for a counter chip to contain two or more *stages* with differing moduli. To take a common example, a divide-by-2 stage and a divide-

by-5 stage that are both present in a single chip can be used to create a decade counter by connecting external pins. The extra stages provide a choice of moduli if they are used individually.

Single and Dual

Counter chips may contain two counters of the same modulus. This is known as a *dual* counter. Dual 4-bit counter chips are common. Each counter can be used separately, or they can be cascaded, in which case the total modulus will be found by multiplying the individual moduli together.

High-State, Low-State, and Three-State

Almost all counters use positive logic where a 1 is represented by a high state and 0 by a low state. Some counters allow an additional output state which has a high impedance and is equivalent to an open circuit. This feature is useful when two or more chips share the same output bus. It is discussed in the entry for **shift registers** in “Three-State Output” on page 126.

Descending Output

Most components only create an ascending count. The output can be converted to a descending count by passing each binary state through an inverter, but this will only work properly if the modulus is equal to the number of states. In a BCD counter, its inverted outputs will count from decimal 15 to decimal 6, not from decimal 9 to decimal 0.

A few counters are available which are designed to create a correct descending count. Other counters are available which allow the user to set the mode to ascending or descending. Examples are the 74x190 or 74x192 (where an acronym for the chip family will be substituted for the letter x).

A descending output is useful in combination with a parallel input, where a user may set an initial value from which the counter will descend

to 0. With suitable logic, this can enable a user-specified delay period.

Programmable Counters

A programmable counter can usually allow a modulus ranging from 2 to more than 10,000. The counter counts down by dividing an initial number repeatedly with a value that is preset with binary inputs. An example is the 4059B chip.

Examples

Many counter specifications date back to the 4000 family of logic chips. Versions of them subsequently became available in the 74xx series, often with the old 4000 part number preceded by 74x (where x is replaced by a designation of the logic family). For example, a version of the 4518B dual BCD chip can be obtained as the 74HC4518. As is the case with all logic chips, this part number will be preceded by letter(s) designating the particular manufacturer, with a suffix added to differentiate minor variants of the chip. The 74xx series has the advantage of higher speed and a greater ability to source or sink current at its output pins.

Most of the original CMOS chips, such as the 4518B, are still available, even in surface-mount versions. These offer the possible advantage of being able to use a higher power supply voltage.

Many counters offer multiple options such as different modulus values that can be selected by external pin connections. Some chips are tolerant of slow clock frequencies; others are not. Most are edge-triggered, but a few are level triggered. Some, such as the 4518B mentioned above, allow a choice of a rising-edge clock input and falling-edge clock input on different pins. For a specific application, it is really necessary to read a variety of datasheets to select the chip that is most suitable.

Values

As is the case with other logic chips, most counters in the through-hole 74xx series are intended

for 5VDC power supply while the older 4000 series may tolerate up to 18VDC. Surface-mount 74xx versions may use voltages as low as 2VDC.

See the section on logic gates in [Chapter 10](#) for a discussion of acceptable high and low logic-input states. On the output side, the 4000 series chips are able to source or sink less than 1mA at 5VDC, but the 74HCxx series can usually manage around 20mA.

A few counters are capable of delivering more power through additional output stages that can drive LEDs. The 4026B decade counter is still being manufactured, capable of powering modest 7-segment displays. The 4033B has the additional option of blanking any leading zeros in a multi-digit display. Other chips that were designed for direct connection to LED numerals have become obsolete as the need for this application has diminished. The 74C925, 74C926, 74C927, and 74C928 are examples. They may still be found from surplus outlets, but should not be specified in new circuit designs.

What Can Go Wrong

The entry that deals with problems affecting shift registers (see [“What Can Go Wrong” on page 128](#)), describes issues which also affect counters. The entry that deals with logic chips (see [“What Can Go Wrong” on page 105](#)), describes problems affecting all types of logic chips. In addition, the potential problems listed below are specific to counters.

Lock-Out

This is the condition which occurs if a counter with a shortened modulus is loaded with a binary state that is out of its range. Consult the datasheet and study its state diagram to determine the most likely outcome if this problem occurs.

Asynchronous Artifacts

Because the flip-flops in an asynchronous (ripple) counter do not change simultaneously, they create very brief false outputs while the ripple

process is taking place. In a 4-bit counter, the binary number 0111 (decimal 7) should be followed by 1000 (decimal 8). However, the rightmost digit (i.e., the least significant bit) will change to a 0 initially, creating 0110 as a momentary binary output (decimal 6). The carry operation will then change the next digit to a 0, creating 0100 (decimal 4). The carry operation continues, changing the next digit to a 0, creating 0000. Finally the operation completes by creating 1000 as the correct output.

These intermediate states on the output pins are often referred to as *glitches*. Because they are extremely brief, they will be undetectable when a counter is used to drive a display. They can cause significant issues, however, if the outputs of the counter are connected with other logic chips.

Another type of asynchronous problem will occur if the clock speed is sufficiently high that a

new pulse is received at the first flip-flop before the ripple of changing states has passed all the way through to the final flip-flop. This will result in a different brief invalid value on the output pins.

Noise

Old TTL-type counters, such as the 74LSxx series, are especially noise-sensitive. Adding a 0.1 μF or 0.047 μF bypass capacitor as close to the power supply pin as possible is recommended. Breadboarding counters of this type may result in errors if a high-frequency clock is used, because conductors such as patch-cords are liable to pick up noise. Modern 74HCxx counters are preferable.

encoder

14.

In this encyclopedia, an **encoder** is a digital chip that converts a decimal-valued input into a binary-coded output.

The term “encoder” may alternatively refer to a **rotational encoder** (also known as a *rotary encoder*) which has a separate entry in Volume 1 of this encyclopedia. The term may also describe a *code hopping encoder*, which is an encryption device used in keyless entry systems for automobiles.

OTHER RELATED COMPONENTS

- **decoder** (see [Chapter 15](#))
- **multiplexer** (see [Chapter 16](#))

What It Does

An encoder is a *logic chip* that receives an input consisting of an active logical state on one of at least four input pins, which have decimal values from 0 upward in increments of 1. The encoder converts the active pin number into a binary value represented by logic states on at least two output pins. This behavior is opposite to that of a **decoder**.

Encoders are identified in terms of their inputs and outputs. For example:

- 4-to-2 encoder (four input pins, two output pins)
- 8-to-3 encoder (eight input pins, three output pins)
- 16-to-4 encoder (sixteen input pins, four output pins)

In the early days of computing, encoders processed interrupts. This application is now rare, and relatively few encoder chips are still being manufactured. However, they are still useful in small devices—for example, if a large number of

inputs must be handled by a microcontroller that has insufficient pins to receive data from each individually.

Schematic Symbol

Like other logic-based components, the encoder does not have a specific schematic symbol and can be represented by a plain rectangle as in [Figure 14-1](#), with inputs on the left and outputs on the right. The bars printed above some of the abbreviations indicate that an input or output is active-low. In this chip, the 74LS148, all inputs and outputs are active-low.

Generally speaking, inputs labeled D0, D1, D2... are used for data input, although they may simply be numbered, with no identifying letter. The encoded outputs are typically identified as Q0, Q1, Q2... or A0, A1, A2... with Q0 or A0 designating the least significant bit in the binary number.

Pins labeled E and GS are explained in the following section.

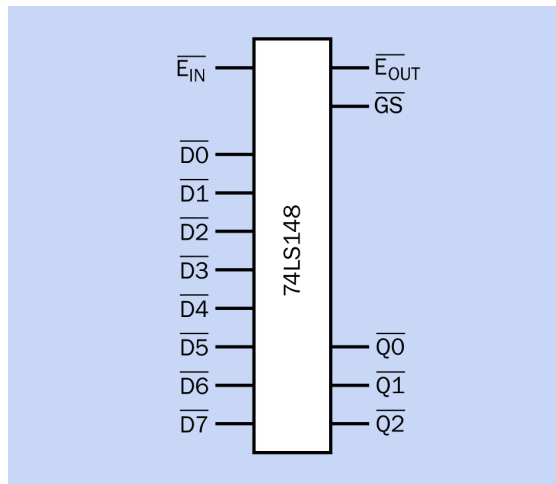


Figure 14-1. While no specific schematic symbol exists for an encoder chip, this style is commonly used. Shown here is a 16-to-4 encoder with active-low inputs and outputs.

Similar Devices

The similarities and differences between encoder, decoder, multiplexer, and demultiplexer can cause confusion.

- In an **encoder**, an active logic state is applied to one of four or more input pins, while the rest remain in an inactive logic state. The input pin number is converted to a binary code which is expressed as a pattern of logic states on two or more output pins.
- In a **decoder**, a binary number is applied as a pattern of logic states on two or more input pins. This value determines which one of four or more output pins will have an active logic state, while the rest remain in an inactive logic state.
- A **multiplexer** can connect a choice of multiple inputs to a single output, for data transfer. The logic state of an enable pin, or a binary number applied as a pattern of logic states to multiple control pins, chooses which input should be connected with the output pin. The alternative term *data selector* evokes the function of this device more clearly.

- An analog multiplexer may allow its inputs and outputs to be swapped, in which case it becomes a **demultiplexer**. It can connect a single input to one of multiple outputs, for data transfer. The logic state of an enable pin, or a binary number applied as a pattern of logic states to multiple control pins, chooses which output should be used. The alternative term *data distributor* evokes the function of this device more clearly.

How It Works

An encoder contains **logic gates**. The internal logic of an 8-to-3 encoder is shown in Figure 14-2, where the darker blue rectangle represents the chip. The switches in this figure are external and are included only to clarify the concept. An open switch is imagined to provide an inactive logic input, while a single closed switch provides an active logic input. (Multiple active inputs can be handled by a *priority encoder*, described below).

Each input switch has a numeric status from 1 to 7. The switch with value 0 does not make an internal connection, because the output from the OR gates is 000 by default.

The logic state of each OR output represents a binary number, weighted with decimal values 1, 2, and 4, as shown at the bottom of the figure. Thus, if switch 5 is pressed, by tracing the connections it is clear that the outputs of OR gates 4 and 1 become active, while the output from gate 2 remains inactive. The values of the active outputs thus sum to 5 decimal.

Figure 14-3 shows the outputs for all possible input states of a 4-to-2 encoder. Figure 14-4 shows the outputs for all possible input states of an 8-to-3 encoder. These diagrams assume that a high logic state is an active logic state, on input or output. This is usually the case.

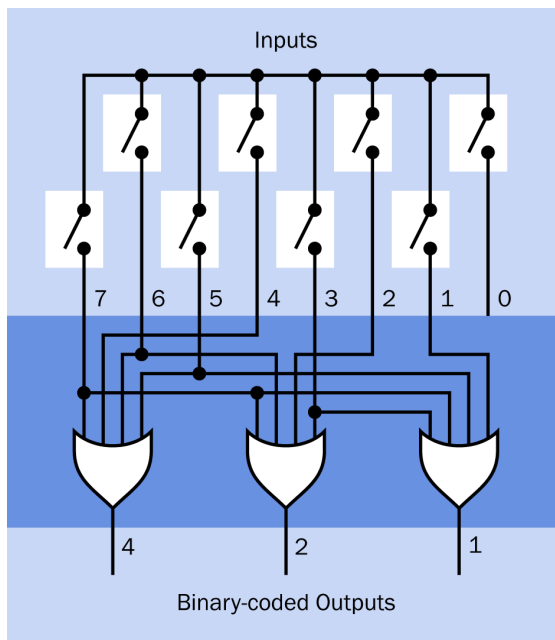


Figure 14-2. A simplified simulation of the internal logic of an 8-to-3 encoder. The dark blue rectangle indicates the space inside the chip. The external switches are included only to clarify the concept. An encoder chip would have an Enable line to create an active output.

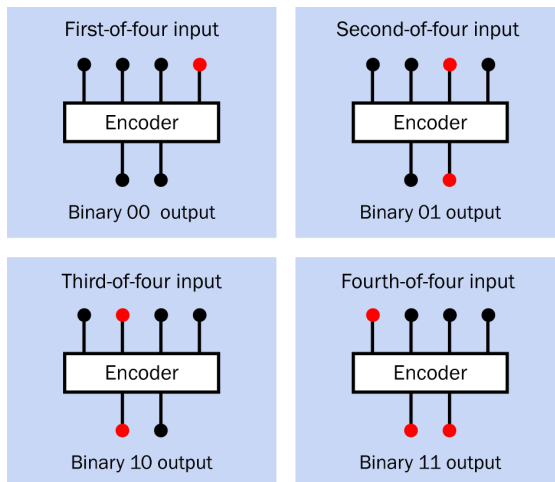


Figure 14-3. The four possible inputs of a 4-to-2 encoder (top of each panel) and the encoded outputs (below).

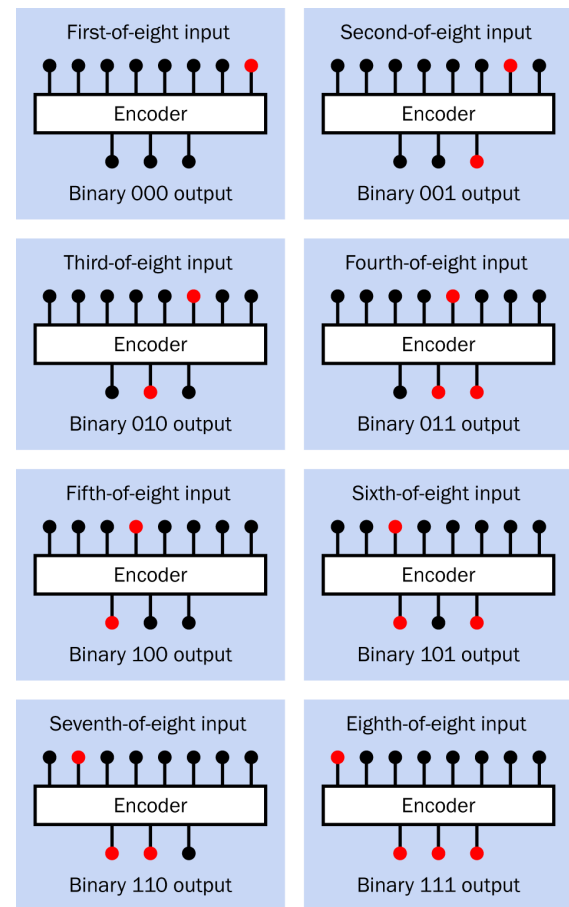


Figure 14-4. The eight possible inputs of an 8-to-3 encoder (at the top of each panel), and the encoded outputs (below). Note that one input of an encoder must always be logic-high. All logic-low inputs are not a valid state.

Unlike ripple counters, where propagation delays can reduce the overall response time of the component, decoders respond within two or three nanoseconds.

Variants

A *simple encoder* assumes that only one input pin can be logically active at a time. A *priority encoder* assigns priority to the highest-value input pin if more than one happens to receive an active input. It ignores any lower-value inputs. An example is the 74LS148, which is an 8-to-3 chip.

A few encoders feature *three-state* outputs (also known as *tri-state*), in which a high-impedance or “floating” output state is available in addition to the usual high and low logic states. The high-impedance state allows multiple chips to share an output bus, as those that are in high-impedance mode appear to be disconnected. This is useful if two or more encoders are cascaded to handle a larger number of inputs.

Values

As is the case with other logic chips, most encoders in the through-hole 74xx series are intended for 5VDC power supply while the older 4000 series may tolerate up to 18VDC. Surface-mount versions may use voltages as low as 2VDC.

See the section on logic gates in [Chapter 10](#) for a discussion of acceptable high and low input states. On the output side, the 4000 series chips are able to source less than 1mA at 5VDC, but the 74HCxx series can manage around 20mA.

How to Use It

Suppose that a microcontroller should respond to an eight-position rotary switch. Because the switch cannot be turned to more than one position at a time, all of its eight contacts can be connected with the inputs on an encoder, which will deliver a 3-bit binary number to three inputs of the microcontroller. Code inside the microcontroller then interprets the pin states.

This is shown in [Figure 14-5](#). Pulldown resistors would be needed on the input pins of the encoder, to prevent them from floating when they are not connected by the rotary switch. They have been omitted from this diagram for simplicity. Debouncing the switch would be handled by the microcontroller.

Other forms of input may be used instead of a rotary switch. For example, the outputs from eight comparators or eight phototransistors could be passed through an encoder.

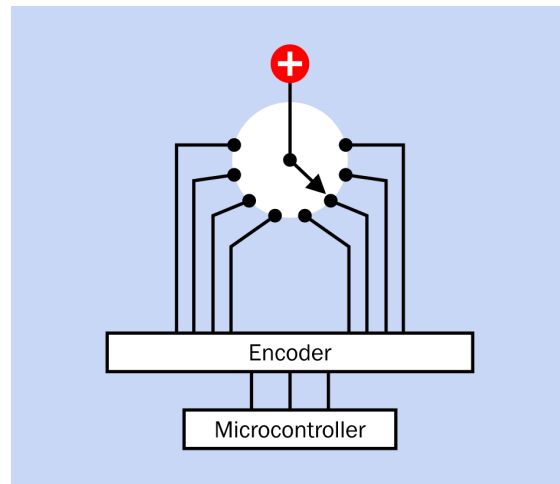


Figure 14-5. Output from an eight-input rotary switch could be connected through an 8-to-3 encoder to provide input to a microcontroller using a reduced number of pins. Pulldown resistors have been omitted for simplicity.

Cascaded Encoders

Encoders are often provided with features to facilitate handling additional inputs via multiple chips. Typically, a second Enable pin is provided, as an output that connects with the Enable input of the preceding chip. This preserves the priority function, so that an input on the second chip prevents any additional input to the first chip from affecting the output. In a datasheet, the enable pins may be labeled E_{IN} and E_{OUT} , or E_I and E_O .

In addition, a GS pin will be included, meaning “Group Select.” It is logically active only when the encoder is enabled and at least one input is active. The GS pin of the most-significant encoder provides an additional binary digit.

The outputs from two encoders can be linked via OR gates, as shown in [Figure 14-6](#), where the lower chip’s GS output provides the most significant bit of a four-bit binary number.

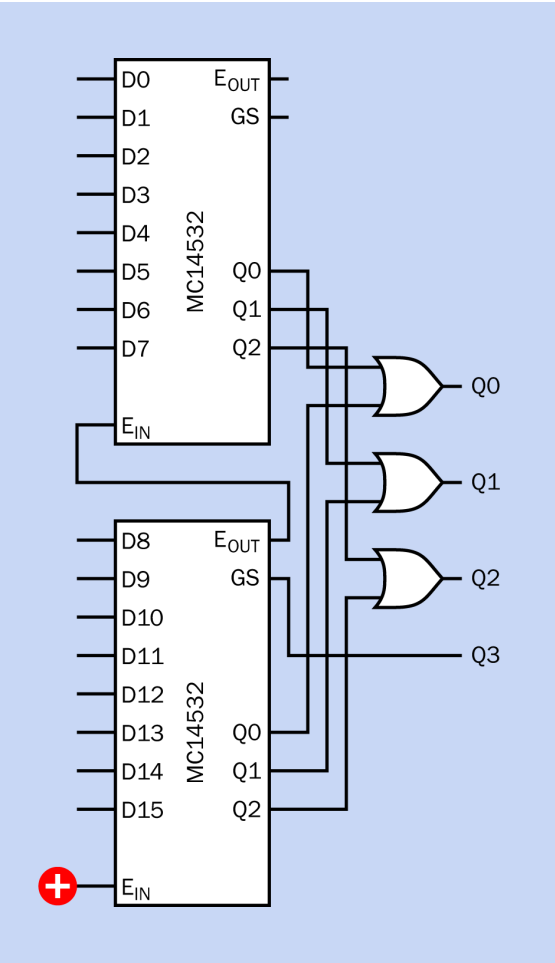


Figure 14-6. Two eight-bit encoders can be cascaded to handle 16 separate inputs. In this example, the encoders use active-high logic.

What Can Go Wrong

Problems that are common to all digital chips are summarized in the section on logic gates in [“What Can Go Wrong” on page 105](#).

See [“What Can Go Wrong” on page 149](#) in the entry describing **decoders** for a list of more specific problems that also afflict encoders.

decoder

15

In this Encyclopedia, a **decoder** is a digital chip that receives a binary-coded input and converts it to a decimal output by applying a logic state to one of a sequence of pins, each of which is assigned an integer value from 0 upward.

The term “decoder” also refers to components and devices that have other functions, such as decoding audio or video formats. These functions are not included here.

OTHER RELATED COMPONENTS

- **encoder** (see [Chapter 14](#))
- **multiplexer** (see [Chapter 16](#))

What it Does

A decoder receives a binary-coded number on two or more input pins. It *decodes* that number and expresses it by activating one of at least four output pins.

The behavior of a decoder with a two-bit binary input is shown in four sequential snapshots in [Figure 15-1](#), where the least significant bit of the input is on the right in each diagram, and the output moves from right to left.

[Figure 15-2](#) shows a similar sequence in a decoder where various values of a three-bit input are decoded to create an eight-pin output.

One sample state of a four-bit decoder is shown in [Figure 15-3](#).

All of these figures assume that a high state represents an active input or output. In a few chips, a low state is used to represent an active output.

Decoders with 2, 3, or 4 input pins are common. To handle a binary input greater than 1111 (decimal 15), decoders can be chained together, as described below.

Manufacturers’ datasheets often describe decoders in terms of their inputs and outputs. Typical examples would include:

- 2-to-4 decoder (two input pins, four output pins)
- 3-to-8 decoder (three input pins, eight output pins)
- 4-to-10 decoder (for converting *binary-coded decimal* to decimal output)
- 4-to-16 decoder (also known as a *hex decoder*).

Input Devices

The input pins of a decoder can be driven by a counter that has a binary-coded output. A decoder can also be driven by a microcontroller, which may have an insufficient number of output pins to control a variety of devices. Two, three, or four of the outputs can be used to represent a binary number which is passed through the decoder to activate the devices one at a time, perhaps with transistors or Darlington arrays introduced to handle the load. This is suggested in [Figure 15-4](#).

A **shift register** can be used for a similar purpose, but often has only one pin for input. This pin must be supplied sequentially with a serial pattern of bits that will match the desired high/low states of the output pins. The relative advantage of this system is that a shift register can generate any pattern of output states. A one-of-many decoder can activate only one output at a time.

LED Driver

A special case is a *seven-segment decoder* designed to drive a seven-segment **LED display** numeral. A binary-coded decimal number on four input pins is converted to a pattern of outputs appropriate for lighting the segments of the display that will form a number from decimal 0 through 9.

Schematic Symbol

Like other logic-based components, the decoder does not have a specific schematic symbol and is represented by a plain rectangle as in [Figure 15-5](#), with inputs on the left and outputs on the right. The bars printed above the E and LE abbreviations (which stand for Enable and Latch Enable, respectively) indicate that they are active-low. In this chip, the 74HC4514, all outputs are active-high, but in a related 4-to-16 decoder, the 74HC4515, all outputs are active-low. In both of these chips, the Enable pin is held low to activate the outputs. The Latch Enable pin freezes the current state of the outputs (i.e., it latches them) when it is held low.

Generally speaking, pins labeled A0, A1, A2... in a datasheet are often the binary inputs (although A, B, C... may be used), with A0 designating the least significant bit. Outputs are usually labeled Y, and are activated in sequence from Y0 when the binary input starts counting upward.

Similar Devices

The similarities and differences between encoder, decoder, multiplexer, and demultiplexer can cause confusion.

- In a **decoder**, a binary number is applied as a pattern of logic states on two or more input pins. This value determines which one of four or more output pins will have an active logic state, while the rest remain in an inactive logic state.
- A **multiplexer** can connect a choice of multiple inputs to a single output, for data transfer. The logic state of an enable pin, or a binary number applied as a pattern of logic states to multiple control pins, chooses which input should be connected with the output pin. The alternative term *data selector* evokes the function of this device more clearly.
- An analog multiplexer may allow its inputs and outputs to be reversed, in which case it becomes a **demultiplexer**. It can connect a single input to one of multiple outputs, for data transfer. The logic state of an enable pin, or a binary number applied as a pattern of logic states to multiple control pins, chooses which output should be used. The alternative term *data distributor* evokes the function of this device more clearly.

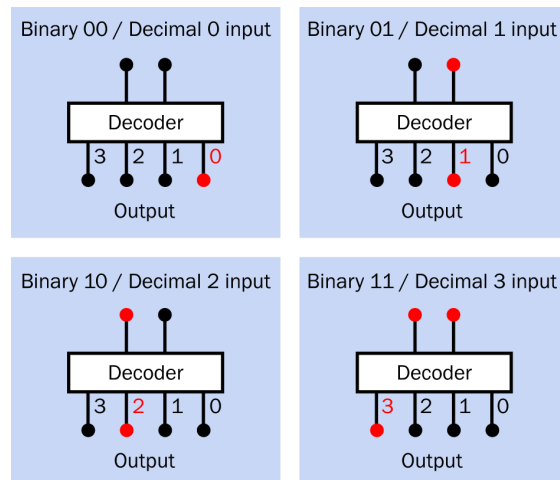


Figure 15-1. A decoder with two input pins can interpret their binary-number representation to create an active logic state on one of four output pins.

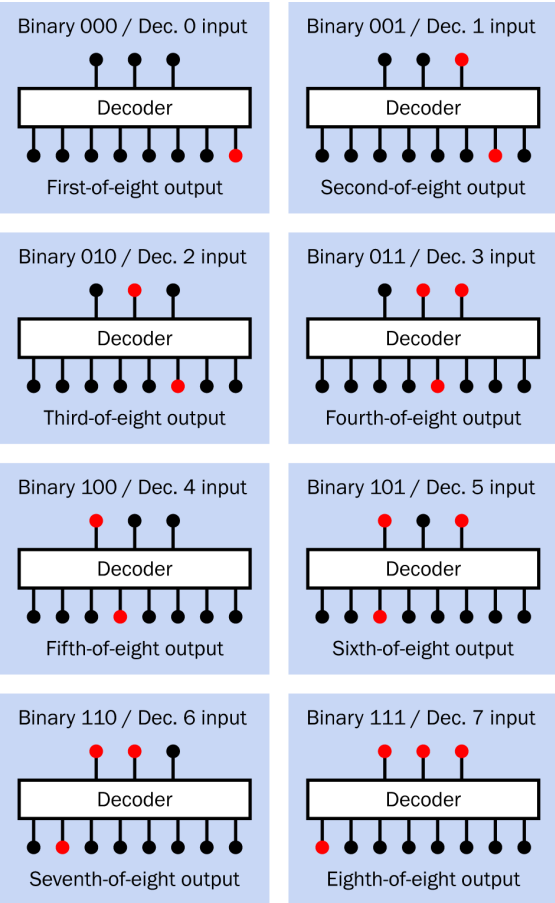


Figure 15-2. A decoder with three input pins can interpret their binary-number representation to create a high logic state on one of eight output pins.

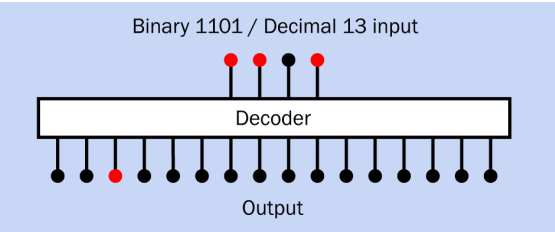


Figure 15-3. A decoder with four input pins can interpret their binary-number representation to create a high logic state on one of 16 output pins. Only one of the 16 possible states is shown here.

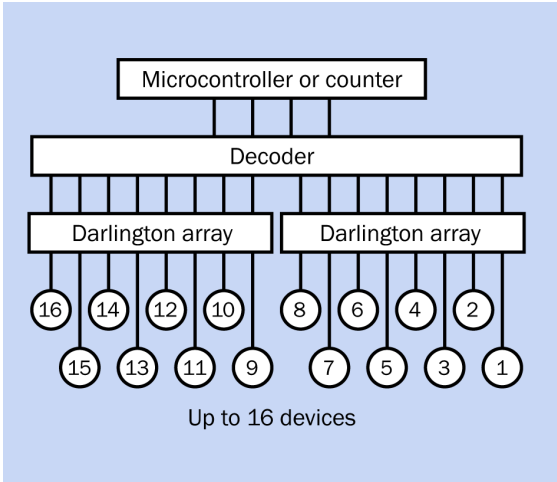


Figure 15-4. Four outputs from a binary counter or microcontroller can be used by a decoder to activate one of up to 16 output devices.

A photograph of a 74HC4514 decoder chip appears in [Figure 15-6](#).

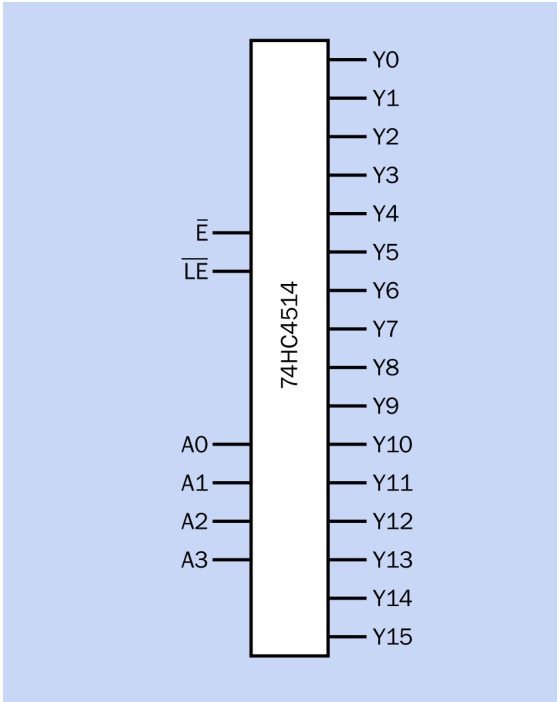


Figure 15-5. While no specific schematic symbol exists for a decoder chip, this style is commonly used. Shown here is a 4-to-16 decoder.



Figure 15-6. The 24-pin 74HC4514 decoder chip processes a 4-bit input and represents it by making one of its 16 output pins active-high.

How It Works

A decoder contains logic gates, each of which is wired to respond to a unique binary pattern of inputs. (In the case of a seven-segment decoder, the internal logic is more complicated.) [Figure 15-7](#) shows the logic of a 2-to-4 decoder. The darker blue area contains the components inside the chip. The external switches are included only to clarify the function of the decoder. An open switch is imagined to provide a low logic input, while each closed switch provides a high logic input.

Unlike ripple counters, where propagation delays can reduce the overall response time of the component, decoders function within two or three nanoseconds.

Variants

Decoder variants have not proliferated with time, and relatively few are available. Most are 3-to-8, 4-to-16, and binary-coded-decimal types.

The 7447 and 74LS47 are seven-segment decoders that have an open-collector output capable of driving a 7-segment display directly. The 7448 is similar but also contains built-in resistors and a capability to blank out leading zeros in a display. However, some suppliers now list the 74LS48 as obsolete. It may be still available from

old stock, but should not be specified in new circuits.

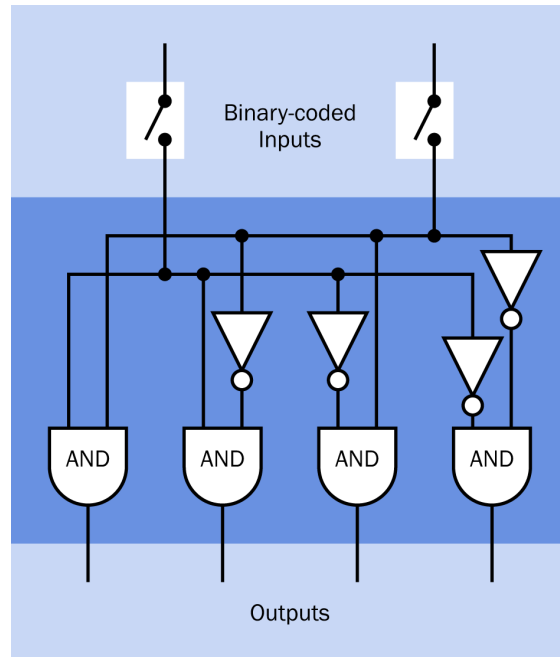


Figure 15-7. A simplified simulation of the logic in a decoder. An actual chip would have an Enable line to activate the output. The dark blue rectangle indicates the space inside the chip.

Although 74LS47 is still being manufactured, and is available in surface-mount as well as through-hole format, a version is not available in the widely used HC family of 74xx chips. Care must be taken to satisfy the input voltage requirements of the 74LS47 when driving it with 74HCxx chips.

Values

As is the case with other logic chips, most decoders in the through-hole 74xx series are intended for 5VDC power supply while the older 4000 series may tolerate up to 18VDC. Surface-mount versions may use voltages as low as 2VDC.

See the section on logic gates in [Chapter 10](#) for a discussion of acceptable high and low input states. On the output side, the 4000 series chips

are able to source or sink less than 1 mA at 5VDC, but the 74HCxx series can manage around 20 mA.

How to Use It

The original applications for decoders in computer circuits have become uncommon, but the chips can still be useful in small appliances and gadgets where multiple outputs are controlled by a **counter** or microcontroller.

Although 16 is usually the maximum number of outputs, some chips are designed to allow expansion. The 74x138 (where a chip family identifier such as LS or HC can be substituted for the letter x) is a 3-to-8 decoder with two logic-low Enable pins and one logic-high Enable. If a value-8 binary line is applied to the low-enable of one chip and the high-enable of another, the first chip will be disabled when the line goes high to indicate that the binary number 1000 has been reached, and the second chip can continue upward from there by sharing the same three less-significant-bit inputs. As many as four chips can be chained in this way.

What Can Go Wrong

Problems that are common to all digital chips are summarized in the section on logic gates in [“What Can Go Wrong” on page 105](#).

Glitches

Although a decoder typically functions faster than a ripple counter, it suffers the same tendency to introduce brief *glitches* in its output. These

are momentary invalid states which occur while processes inside the chip that are slightly slower are catching up with other processes that reach completion slightly faster. A brief *settling time* is necessary to ensure that the output is stable and valid. This will be irrelevant when powering a device such as an **LED indicator**, which will not display such brief transients. The problem may be more important if the output from the decoder is used as an input to other logic chips.

If the input to a decoder is derived from a ripple counter, the input may also contain glitches, which can cause erroneous outputs from the decoder. It is better to use a synchronous counter on the input side of a decoder.

Unhelpful Classification

Online parts suppliers tend to list decoders under the same category heading as encoders, multiplexers, and demultiplexers, making it difficult to find what you want. Under this broad subject heading (which will include thousands of chips), if you search by selecting the number of inputs relative to the number of outputs that you have in mind, this will narrow the search considerably.

Active-Low and Active-High

Chips with identical appearance and similar part numbers may have outputs that are either active-low or active-high. Some may offer a latch-enable pin, while others have enable pins that must be pulled low or forced high to produce an output. Accidental chip substitution is a common cause of confusion.

multiplexer

16

May be abbreviated as a *mux* (this term is sometimes printed all in caps), and may be referred to alternatively as a *data selector*. Some sources maintain that a multiplexer has no more than two channels, whereas a data selector has more, but there is no consensus on this, and datasheets continue to use the term “multiplexer” predominantly.

Analog multiplexers are usually bidirectional, and thus will function equally well as *demultiplexers*. Consequently, this encyclopedia does not contain a separate entry for demultiplexers.

OTHER RELATED COMPONENTS

- **encoder** (see [Chapter 14](#))
- **decoder** (see [Chapter 15](#))

What It Does

A multiplexer can select one of two or more input pins, and connect it internally with an output pin. Although it is an entirely solid-state device, it behaves as if it contains a **rotary switch** in series with a SPST switch, as shown in [Figure 16-1](#). A binary code applied to one or more Select pins chooses the input, and an Enable pin establishes the connection with the output. The Select and Enable functions are processed via an internal section referred to as a *decoder*, not to be confused with a **decoder** chip, which has its own entry in this encyclopedia.

All multiplexers are digitally controlled devices, but may be described as either *digital* or *analog* depending how they process the input signal. A digital multiplexer creates an output that is adjusted to logic-high or logic-low within the limits of its logic family. An analog multiplexer does not impose any processing on the voltage, and passes along any fluctuations. Thus, it can be used with alternating current.

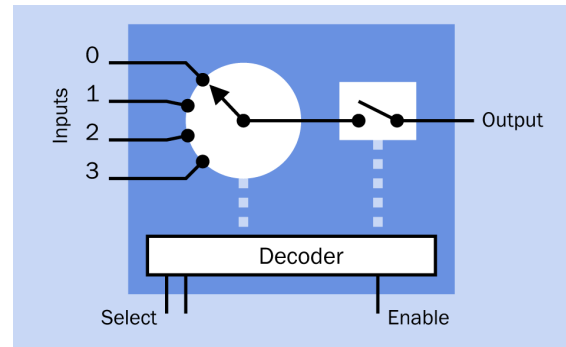


Figure 16-1. A multiplexer functions as if it contains a rotary switch. The switch position is determined by a binary number applied to external Select pins. The internal connection is completed by applying a signal to an Enable pin.

Because an analog multiplexer merely switches a flow of current, it can be *bidirectional*; in other words, it can function as a *demultiplexer*, in which case the input is applied to the pole of the (imaginary) internal switch and outputs are taken from the terminals.

Differential Multiplexer

A *differential* multiplexer contains multiple switches that are differentiated from one another (i.e., they are electrically isolated, although they are controlled by the same set of select pins). A differential multiplexer is conceptually similar to a rotary switch with two or more decks controlled by a single shaft. See [Figure 16-2](#).

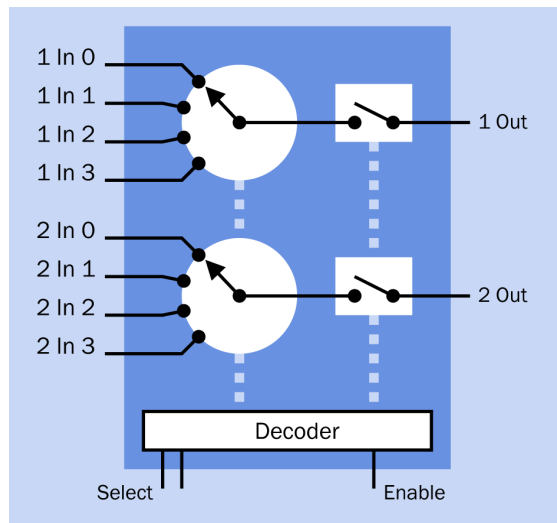


Figure 16-2. A differential multiplexer contains two or more electronic switches that are differentiated from one another, similarly to the decks on a rotary switch. Although the channels into each switch are typically numbered from 0 upward, the switches are numbered from 1 upward.

A bidirectional dual 4-channel differential analog multiplexer is shown in [Figure 16-3](#).

Modern multiplexers are often found switching high-frequency data streams in audio, telecommunications, or video applications.

Similar Devices

The similarities and differences between multiplexer, demultiplexer, encoder, and decoder can cause confusion:

- A **multiplexer** can connect a choice of multiple inputs to a single output, for data transfer. The logic state of an enable pin, or a binary number applied as a pattern of logic

states to multiple control pins, chooses which input should be connected with the output pin. The alternative term *data selector* evokes the function of this device more clearly.

- An *analog* multiplexer may allow its inputs and outputs to be reversed, allowing it to become a **demultiplexer**, connecting a single input to one of multiple outputs, for data transfer. The logic state of an enable pin, or a binary number applied as a pattern of logic states to multiple control pins, chooses which output should be used. The alternative term *data distributor* evokes the function of this device more clearly.

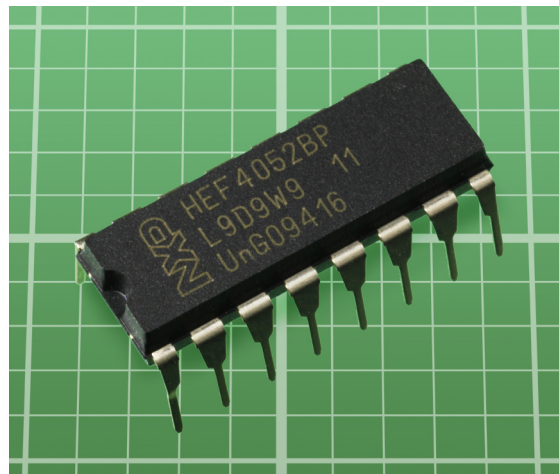


Figure 16-3. This CMOS chip contains two four-channel differential analog multiplexers.

- In an **encoder**, an active logic state is applied to one of four or more input pins, while the rest remain in an inactive logic state. The input pin number is converted to a binary code which is expressed as a pattern of logic states on two or more output pins.
- In a **decoder**, a binary number is applied as a pattern of logic states on two or more input pins. This value determines which one of four or more output pins will have an active logic state, while the rest remain in an inactive

logic state. A digital multiplexer does not allow reversal of its inputs and outputs, but a decoder functions as if it were a digital demultiplexer.

How It Works

The multiple inputs to a multiplexer are referred to as *channels*. Almost always, the number of channels is 1, 2, 4, 8, or 16. A 1-channel component is only capable of “on” or “off” modes and functions similarly to a SPST switch.

If there are more than two channels, a binary number will determine which channel is connected internally. The number of channels is usually the maximum that can be identified by the number of select pins, so that 2 pins will control 4 channels, 3 pins will control 8 channels, and 4 pins (the usual maximum) will control 16 channels.

In multiplexers with three or more channels, an enable pin is usually still present to activate or deactivate all the channels simultaneously. The enable feature may be described alternatively as a *strobe*, or may have an inverse function as an *inhibit* pin.

Although a rotary switch is helpful in conceptualizing the function of a multiplexer, a more common representation (sometimes in datasheets) is an array of SPST switches, each of which can be opened or closed by the decoder circuit. A typical example, depicting a dual differential multiplexer, is shown in Figure 16-4. Note that the internal decoder can only close one switch in each channel at a time.

The switch analogy is appropriate in that when an output from a multiplexer is not connected internally (i.e., its switch is “open”) it is effectively an open circuit. However, some multiplexers contain *pullup resistors* to give each output a defined state. This can be an important factor in determining whether the multiplexer is suitable for a particular application.

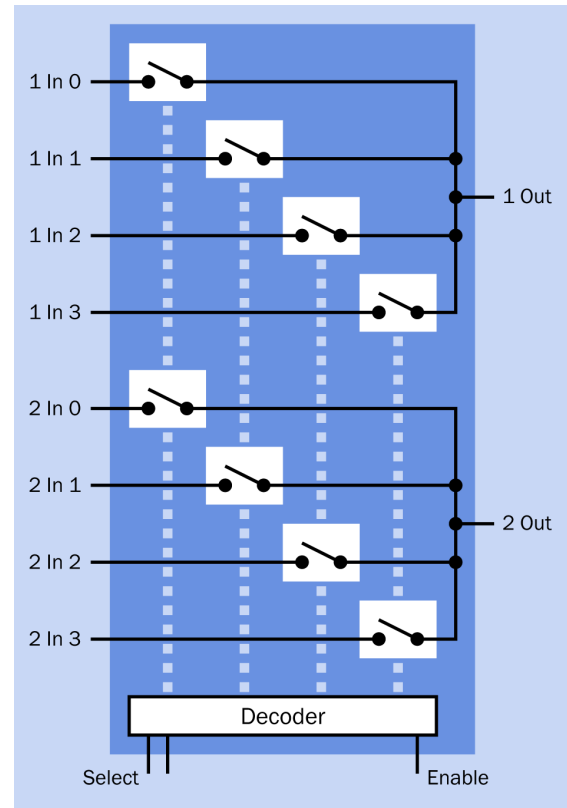


Figure 16-4. The internal function of a dual multiplexer is commonly represented as a network of SPST switches, each of which is controlled by decoder logic.

A digital multiplexer actually contains a network of logic gates, shown in simplified form in Figure 16-5.

A demultiplexer has internal logic shown in simplified form in Figure 16-6.

Schematic Symbol

In a schematic, a multiplexer and demultiplexer may be represented by a trapezoid with its longer vertical side oriented toward the larger number of connections. This is shown in Figure 16-7. However, this symbol is falling into disuse.

More often, as is the case with most logic components, a multiplexer or demultiplexer is represented by a rectangle with inputs on the left and

outputs on the right, as shown in [Figure 16-8](#). The distinction between inputs and outputs is problematic, however, in an analog multiplexer which will allow data flow to be reversed.

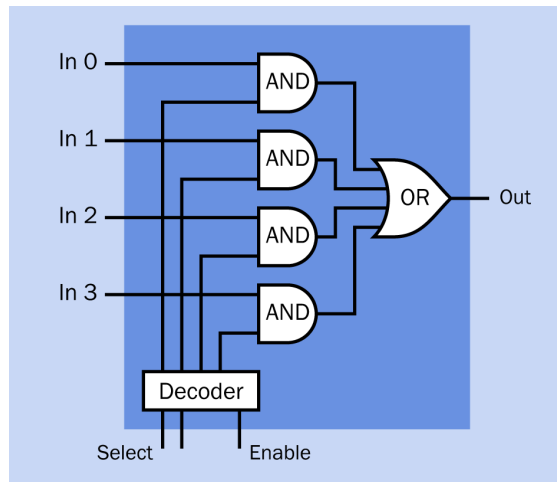


Figure 16-5. A simplified representation of the logic gates in a digital multiplexer.

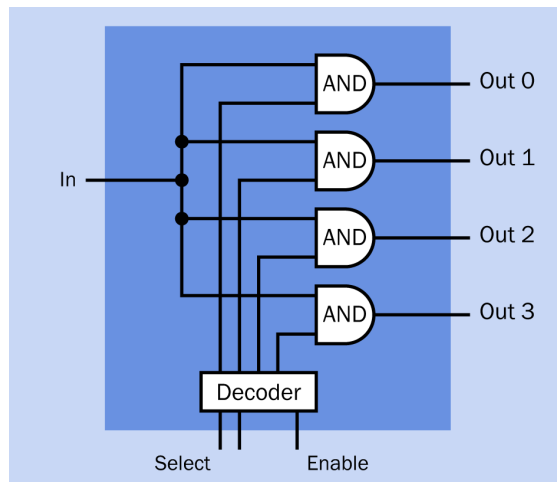


Figure 16-6. A simplified representation of the logic gates in a digital demultiplexer.

Pin Identifiers

The lack of standardization in the identification of pin functions is perhaps more extreme in the case of multiplexers than for other types of logic chips.

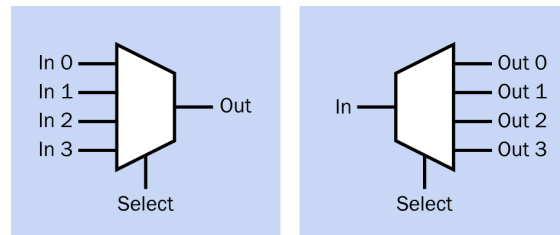


Figure 16-7. The traditional symbol for a multiplexer (left) and demultiplexer (right). The trapezoid is oriented with its longer vertical side facing the larger number of connections. This symbol is falling into disuse.

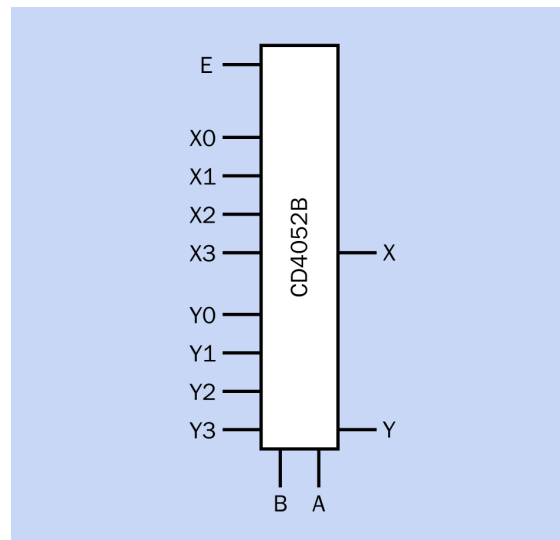


Figure 16-8. A simple rectangle is most often used as a schematic symbol for a multiplexer, but the abbreviations assigned to pin functions are not standardized. See text for details.

An output enable pin will be shown as E or EN, or occasionally OE. It may alternatively be described as an inhibit pin, labeled INH, or sometimes will be called a strobe. The function is the same in each case: one of its logic states will enable the internal switches, while its other logic state will prevent any internal switches from closing.

Switch inputs may be labeled S0, S1, S2... or X0, X1, X2... or may simply be numbered, almost always counting up from 0. Where two or more sets of switches coexist in one package, each set of

inputs may be distinguished from the others by preceding each identifier with a numeral or letter to designate the switch, as in 1S0, 1S1, 1S2... or 1X0, 1X1, 1X2... (Switches are generally numbered from 1 upward, even though their inputs are numbered from 0 upward.) Outputs may be identified using the same coding scheme as inputs, bearing in mind that the inputs and outputs of an analog multiplexer usually are interchangeable. Some manufacturers, however, prefer to identify each multiplexer output by preceding it with letter Y. Alternatively, Z1, Z2, Z3... may identify the outputs from switches 1, 2, 3... Fortunately, datasheets usually include some kind of key to this grab-bag of abbreviations.

Control pins are often identified as A, B, C... with letter A representing the least significant bit in the binary number that is applied to the pins.

Voltages can be confusing in multiplexers. Components intended for use with digital inputs are straightforward enough, as the supply voltage will be identified as V_{CC} and is typically 5VDC for through-hole packages (often lower for surface-mount), while negative ground is assumed to be 0VDC. However, where a multiplexer may be used with AC inputs in which the voltage varies above and below 0V, supply voltages above and below 0VDC are also possible—such as +7.5VDC and −7.5VDC, to take a random example. Three power-supply pins may be provided for this purpose. The positive supply will usually be identified as V_{DD} (the D refers to the Drain in the internal MOSFETs). A V_{EE} pin may be at 0VDC or at a negative value equal and opposite to V_{DD} . The E in this abbreviation is derived from Emitter voltage, even though the component may not contain a bipolar transistor with an emitter. Customarily, a V_{SS} pin (the S being derived from the Source in the internal MOSFETs) will be at 0VDC, and other voltages will be measured above and below this baseline. This ground pin may alternatively be labeled GND.

As is customary in logic chips, low-active control pins will have a bar printed above their identifier,

or an apostrophe will be placed after an identifier if the font does not permit printing the bar. Alternatively, low-active pins may be represented by showing a small circle, properly referred to as a *bubble*, at the input or output point of the symbol for the multiplexer. Note that analog inputs and outputs are neither high-active nor low-active; they merely pass voltages through.

Variants

Most multiplexers are “break before make” devices, where one input is disconnected before the next input is connected. However, some exceptions exist, and datasheets should be checked for this. It can be a significant issue, because make-before-break switching will briefly connect external devices with each other, through the chip.

Many multiplexers can tolerate control voltages above the usual high value in a logic circuit—as high as 15VDC in some cases. The voltage that is switched by the multiplexer may be the same as the control voltage, or may be higher.

Some analog multiplexers have *overvoltage protection* that allows them to withstand input voltages that are twice or three times the recommended maximum.

Datasheets may mention “internal address decoding,” meaning that the binary number input, specifying a channel to be switched, is decoded inside the chip. In fact, virtually all multiplexers now have on-chip address decoding, and this feature should be assumed to exist, regardless of whether it is mentioned.

Values

The voltage to be switched will usually be referred to as the *input voltage*, V_{IN} .

An analog multiplexer should not be subjected to current exceeding the value that it is designed to switch. This is known as the maximum *channel current*. A typical value would be 10mA, although many modern surface-mount components are designed for currents in the microamp range.

The **on-resistance** is the resistance imposed by the analog multiplexer on the signal flowing through it. While modern, specialized analog multiplexers may have an on-resistance as low as 5Ω , these are relatively unusual. An on-resistance of 100Ω to 200Ω is more common. This value will vary within a component depending on the power supply voltage and the voltage being switched. It will increase slightly as V_{IN} deviates above (or below) $0V$, will increase substantially for lower values of supply voltage, and will increase significantly with temperature.

The curves in Figure 16-9 show on-resistance of an analog multiplexer varying with input voltage, with three different power supplies: plus-and-minus 2.5VDC (described in the graph as a “spread” of 5VDC), plus-and-minus 5VDC (a “spread” of 10VDC), and plus-and-minus 7.5VDC (a “spread” of 15VDC). These curves were derived from a datasheet for the MC14067B analog multiplexer; curves for other chips will be different, although the basic principles remain the same.

Switching time is an important consideration in high-speed applications. The “on” and “off” times specified in a datasheet (often as t_{ON} and t_{OFF}) are a function of the propagation delay from the control input to the toggling of the switch, and are generally measured from the halfway point of the rising or falling edge of the control input, to the 90% point of the output signal level.

Leakage current is the small amount of current (often measured in picoamperes) that the solid-state switch will pass when it is in its “off” state. This should be insignificant except when very high-impedance loads are used.

Separate switches inside a multiplexer may have characteristics that differ slightly from one another. Differences in on-resistance between adjacent switches can be important when switching parallel analog signals. A datasheet should mention the extent to which switches have matched characteristics, and may define the maximum deviation from one another using the abbreviation R_{ON} even though this same term

may be used, confusingly, to denote the on-resistance of each individual switch.

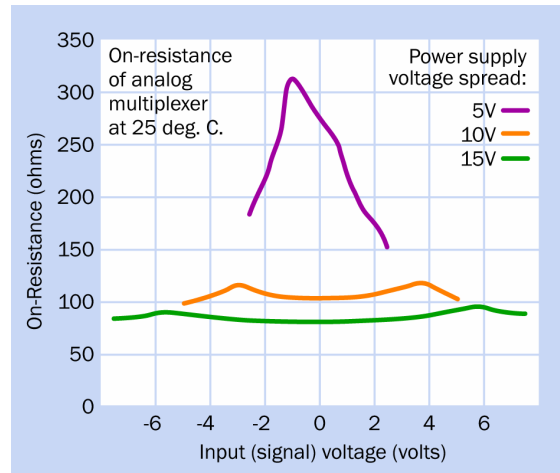


Figure 16-9. Variations in on-resistance in an analog multiplexer. Each voltage “spread” is the difference between positive supply voltage and an equal-and-opposite negative ground voltage. Thus a “spread” of 10VDC means plus and minus voltages of 5VDC. (Curves derived from On Semiconductor datasheet for MC14067B analog multiplexer.)

How to Use It

A multiplexer may be used as a simple switch to choose one of multiple inputs, such as a choice of input jacks on a stereo system. A dual differential multiplexer is useful in this application, as it can use a single select signal to switch two signal paths simultaneously.

A multiplexer can also be used as a digital volume control by switching an audio signal among a variety of resistances, similar to a **digital potentiometer**. In this application, the possible presence of pullup resistors inside the multiplexer must be considered.

Where a microcontroller must monitor a large number of inputs (for example, a range of temperature sensors or motion sensors), a multiplexer can reduce the number of input pins required. Its data-select pins will be cycled through all the possible binary states by the microcontroller, to select each data input in turn, while its single-

wire output will carry the analog data to a separate pin on the microcontroller which performs an analog-digital conversion.

Conversely, a demultiplexer (i.e., an analog multiplexer such as the 4067B chip which can be used in demultiplexer mode) can be used by a microcontroller to switch multiple components on and off. Four outputs from the microcontroller can connect with the control pins of a 16-channel demultiplexer, counting from binary 0000 through binary 1111 to select output pins 0 through 15. After selecting each pin, the microcontroller can send a high or low pulse through it. The process then repeats. (A **decoder** can be used in the same way.)

Other Application Notes

Multiplexers may be *cascaded* to increase the inputs-to-outputs ratio.

Modern multiplexers are found on computer boards where they choose among video output ports, or as PCI express channel switches.

A multiplexer may be used as a parallel-to-serial converter, as it samples multiple channels and converts them into a serial data stream.

In telecommunications, a multiplexer can sample voice signals from multiple separate inputs and combine them into a digital stream that can be transmitted at a faster bit rate over a single channel. However, this application goes far beyond the simple uses for multiplexers described here.

What Can Go Wrong

Problems that are common to all digital chips are summarized in the section on logic gates (see “What Can Go Wrong” on page 105).

Pullup Resistors

While they are often necessary to prevent connections from floating, pullup resistors built into

a multiplexer may have unexpected consequences if the user is unaware of them.

Break Before Make

For most applications, it is desirable for each internal solid-state switch to break one connection before making a new one. This avoids the possibility of separate external components being briefly connected with each other through the multiplexer. Datasheets should be checked to verify that a multiplexer functions in break-before-make mode. If it doesn't, the enable pin can be used momentarily to disable all connections before a new connection is established.

Signal Distortion

Where a multiplexer is passing analog signals, signal distortion can result if the on-resistance of multiple internal switches varies significantly at different voltages. A datasheet for an analog multiplexer should usually include a graph showing on-resistance over the full signal range. The flatter the graph is, the less distortion the component will create. This is often described in datasheets as R_{ON} Flatness.

Limits of CMOS Switching

Although most multiplexers are built around CMOS transistors, their switching speed may be insufficient for video signals, and their on-resistance may vary enough to introduce distortion. Multiplexers are available with complementary bipolar switching for very high-speed applications. They impose some penalties in cost and power consumption.

Transients

Switch capacitance inside a multiplexer can cause transients in the output when the switch changes state. An allowance for *settling time* may be necessary. This will be additional to the switching speed claimed by the datasheet.

LCD

17

The full term *liquid-crystal display* is seldom used. Its acronym, **LCD**, is much more common. Sometimes the redundant combination **LCD display** is found. All three terms refer to the same device. In this encyclopedia, the first two words in *liquid-crystal display* are hyphenated because they are an adjectival phrase. Other sources often omit the hyphen.

The acronym **LED** (for **light-emitting diode**) is easily confused with **LCD**. While both devices display information, their mode of action is completely different.

OTHER RELATED COMPONENTS

- **LED display** (see [Chapter 24](#))

What It Does

An LCD presents information on a small display panel or screen by using one or more segments that change their appearance in response to an AC voltage. The display may contain alphanumeric characters and/or symbols, icons, dots, or pixels in a *bitmap*.

Because of its very low power consumption, a basic monochrome LCD is often used to display numerals in battery-powered devices such as digital watches and calculators. A small liquid-crystal display of this type is shown in [Figure 17-1](#).

Color-enabled, backlit LCDs are now frequently used in almost all forms of video displays, including those in cellular telephones, computer monitors, game-playing devices, TV screens, and aircraft cockpit displays.

How It Works

Light consists of electromagnetic waves that possess an electric field and a magnetic field. The fields are perpendicular to each other and to the direction in which the light is traveling, but the field polarities are randomly mixed in most visi-

ble radiation. This type of light is referred to as *incoherent*.

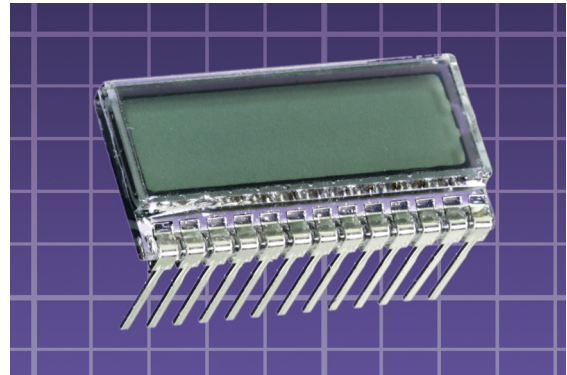


Figure 17-1. A small, basic monochrome LCD.

[Figure 17-2](#) shows a simplified view of an LCD that uses a backlight. Incoherent light emerges from the backlight panel (A) and enters a vertical polarizing filter (B) that limits the electric field vector. The polarized light then enters a liquid crystal (C) which is a liquid composed of molecules organized in a regular helical structure that rotates the polarity by 90 degrees when no voltage is applied to it. The light now passes through

a horizontal polarizing filter (D) and is visible to the user.

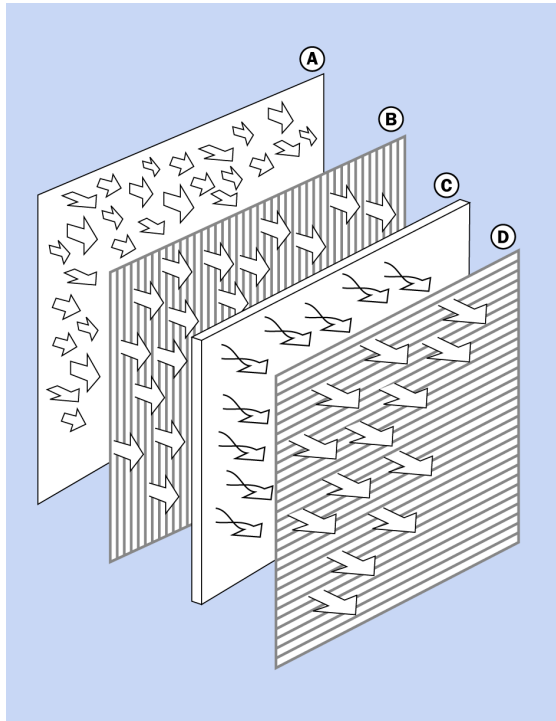


Figure 17-2. The combination of two polarizers and a liquid crystal appears transparent when voltage is not applied. See text for details.

- A liquid crystal itself does not emit light. It can only modify light that passes through it.

Figure 17-3 shows what happens when voltage is applied to the liquid crystal via transparent electrodes (not included in the figure). The molecules reorganize themselves in response to the electric potential and allow light to pass without changing its polarity. Consequently, the vertically polarized light is now blocked by the front, horizontally polarized filter, and the display becomes dark.

A liquid crystal contains ionic compounds that will be attracted to the electrodes if a DC voltage is applied for a significant period of time. This can degrade the display permanently. Therefore, AC

voltage must be used. An AC frequency of 50Hz to 100Hz is common.

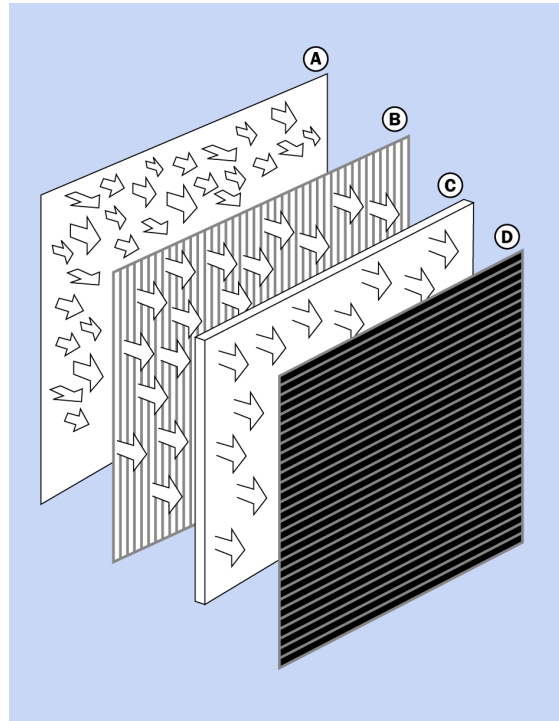


Figure 17-3. The LCD appears dark when voltage is applied. See text for details.

Variants

A *transmissive LCD* requires a backlight to be visible, and is the type illustrated in Figure 17-2. In its simplest form, it is a monochrome device, but is often enhanced to display full color by adding red, green, and blue filters. Alternatively, instead of a white backlight, an array of pixel-sized red, green, and blue **LEDs** may be used, in which case filters are unnecessary.

Backlit color LCDs have displaced *cathode-ray tubes*, which used to be the default system in almost all video monitors and TVs. LCDs are not only cheaper but can be fabricated in larger sizes. They do not suffer from *burn in*, where a persistent unchanging image creates a permanent scar in the phosphors on the inside of a tube. However, large LCDs may suffer from dead pixels or

stuck pixels as manufacturing defects. Different manufacturers and vendors have varying policies regarding the maximum acceptable number of pixel defects.

In a *reflective LCD*, the structure is basically the same as that shown in [Figure 17-2](#) except that a reflective surface is substituted for the backlight. Ambient light enters from the front of the display, and is either blocked by the liquid crystal in combination with the polarizing filters, or is allowed to reach the reflective surface at the rear, from which it reflects back through the liquid crystal to the eye of the user. This type of display is very easily readable in a bright environment, but will be difficult to see in dim conditions and will be invisible in darkness. Therefore, it may be augmented with a user-activated light source mounted at the side of the display.

A *transreflective LCD* contains a translucent rear polarizer that will reflect some ambient light, and is also transparent to enable a backlight. While this type of LCD is not as bright as a reflective LCD and has less contrast, it is more versatile and can be more energy efficient, as the backlight can be switched off automatically when ambient light is bright enough to make the display visible.

Active and Passive Types

An *active matrix* LCD adds a matrix of thin-film transistors to the basic liquid-crystal array, to store the state of each segment or pixel actively while the energizing AC voltage transitions from positive to negative. This enables a brighter, sharper display as *crosstalk* between adjacent pixels is reduced. Because *thin-film transistors* are used, this is often described as a *TFT* display; but the term is interchangeable with *active matrix*.

A *passive matrix* LCD is cheaper to fabricate but responds sluggishly in large displays and is not so well suited to fine gradations in intensity. This type of component is used primarily in simple monochrome displays lacking intermediate shades of gray.

Crystal Types

Twisted Nematic (TN) are the cheapest, simplest type of LCD, allowing only a small viewing angle and average contrast. The appearance is limited to black on gray. The response rate is relatively slow.

Super Twisted Nematic (STN) displays were developed in the 1980s for passive LCDs, enabling better detail, wider view angle, and a faster response. The natural appearance is dark violet or black on green, or dark blue on silver-gray.

Film-compensated Super Twisted Nematic (FSTN) uses an extra coating of film that enables a pure black on white display.

Double Super Twisted Nematic provides further enhancement of contrast and response times, and automatic contrast compensation in response to ambient temperature. The appearance is black on white. This display requires backlighting.

Color Super Twisted Nematic (CSTN) is an STN display with filters added for full color reproduction.

Seven-Segment Displays

The earliest monochrome LCDs in devices such as watches and calculators used seven segments to display each numeral from 0 through 9. This type of LCD is still used in low-cost applications. A separate control line, or electrode, connects to each segment, while a *backplane* is shared by all the segments, connecting with a *common* pin to complete the circuit.

[Figure 17-4](#) shows a typical seven-segment display. The lowercase letters *a* through *g* that identify each segment are universally used in data-sheets. The decimal point, customarily referred to as “dp,” may be omitted from some displays. The array of segments is slanted forward to enable more acceptable representation of the diagonal stroke in numeral 7.

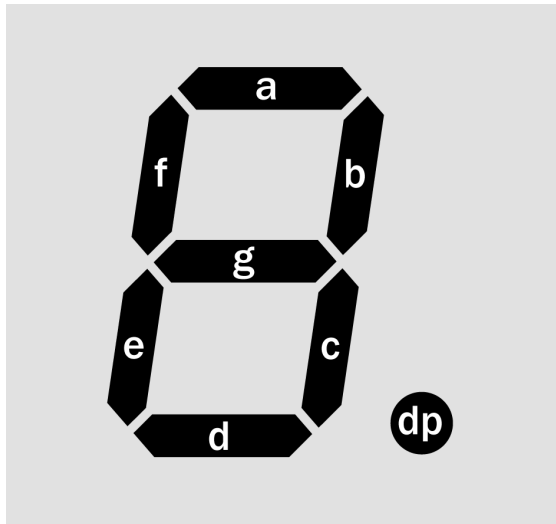


Figure 17-4. Basic numeric display format for LCD numeric displays (the same layout is used with LEDs). To identify each segment, lowercase letters are universally used.

Seven-segment displays are not elegant but are functional and are reasonably easy to read, as shown in Figure 17-5. Letters A, B, C, D, E, and F (displayed as A, b, c, d, E, F because of the restrictions imposed by the small number of segments) may be added to enable display of hexadecimal values.

In appliances such as microwave ovens, very basic text messages can be displayed to the user within the limitations of 7-segment displays, as suggested in Figure 17-6.

The advantage of this system is low cost, as 7-segment displays are cheap to fabricate, entail the fewest connections, and require minimal decoding to create each alphanumeric character. However, numbers 0, 1, and 5 cannot be distinguished from letters O, l, and S, while letters containing diagonal strokes, such as K, M, N, W, X, and Z, cannot be displayed at all.



Figure 17-5. Numerals and the first six letters of the alphabet created with a 7-segment display.

Additional Segments

Alphanumeric LCDs were developed using 14 or 16 segments to enable better representation of letters of the alphabet. Sometimes these displays were slanted forward, like the 7-segment displays, perhaps because the style had become familiar, even though the addition of diagonal segments made it unnecessary. In other cases, the 14 or 16 segments were arrayed in a rectangle. See Figure 17-7.

The same words represented in Figure 17-6 are shown in Figure 17-8, using 16-segment LCDs. Clearly, the advantage gained by enabling diagonal strokes entailed the disadvantage of larger gaps in the letters, making them ugly and difficult to read.

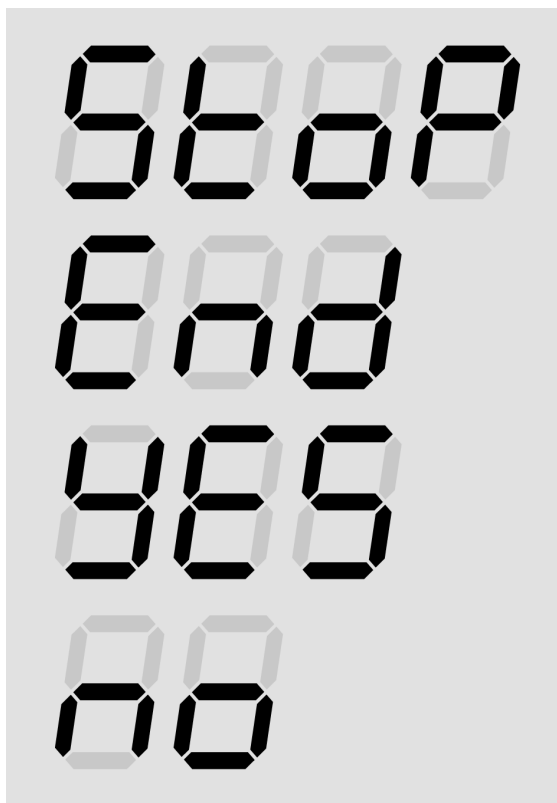


Figure 17-6. Basic text messages can be generated with 7-segment displays, although they cannot contain alphabetical letters that use diagonal strokes.

A full character set using 16-segment LCDs is shown in [Figure 17-9](#). This conforms partially with the ASCII coding system, in which each character has an identifying numeric code ranging from 20 hexadecimal for a letter-space to 7A hexadecimal for letter z (although this character set does not attempt to represent lowercase letters differently from uppercase). The ASCII acronym stands for *American Standard Code for Information Interchange*.

Because backlit LCDs had become common by the time 16-segment displays were introduced, the characters were often displayed in light-on-dark or “negative” format, as suggested in this figure. LEDs, of course, have always used the light-on-dark format, as an LED is a light-emitting component.

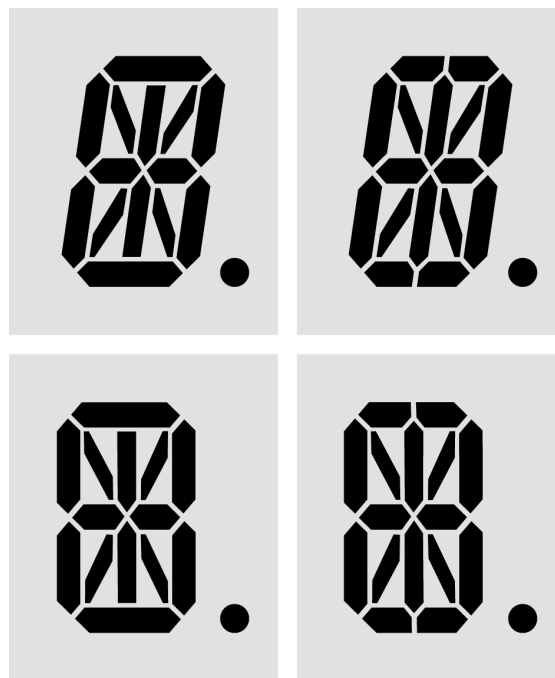


Figure 17-7. LCDs using 14 segments (left) and 16 segments (right) were introduced to represent a full alphabet in addition to numerals. Sometimes these displays were slanted forward, like the previous 7-segment type, even though this was no longer necessary to represent the number 7.

Dot-Matrix Displays

The 16-segment displays were never widely popular, and the declining cost of microprocessors, LCD fabrication, and ROM storage made it economic to produce displays using the more easily legible 5x7 dot-matrix alphabet that had been common among early microcomputers. [Figure 17-10](#) shows a dot-matrix character set that is typical of many LCDs.

Because the original ASCII codes were not standardized below 20 hexadecimal or above 7A hexadecimal, manufacturers have represented a variety of foreign-language characters, Greek letters, Japanese characters, accented letters, or symbols using codes 00 through 1F and 7B through FF. The lower codes are often left blank, allowing user installation of custom symbols. Codes 00 through 0F are often reserved for control functions, such as a command to start a new

line of text. There is no standardization in this area, and the user must examine a datasheet for guidance.

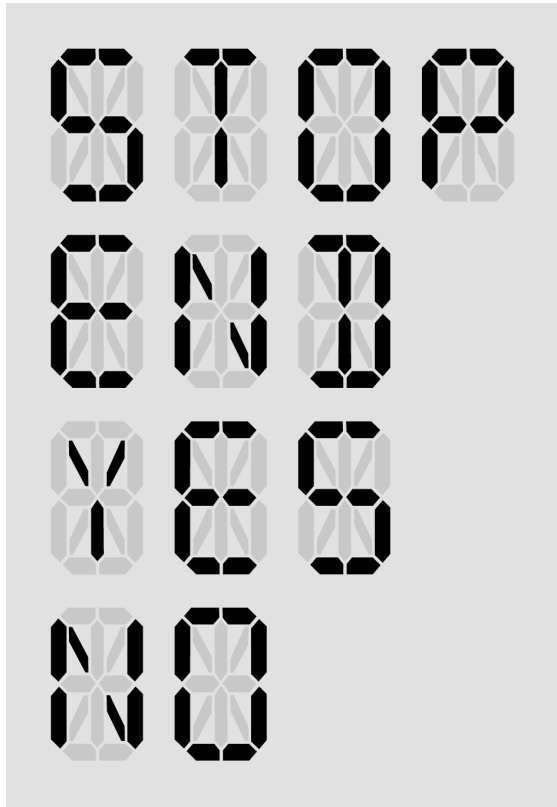


Figure 17-8. The same text messages shown previously using 7-segment LCDs are shown here using 16-segment displays.

Dot-matrix LCDs are usually packaged in arrays consisting of eight or more columns and two or more rows of characters. The number of columns is always stated before the number of rows, so that a typical 8 x 2 display contains eight alphanumeric characters in two horizontal rows. An array of characters is properly referred to as a *display module*, but may be described, confusingly, as a *display*, even though a single seven-segment LCD is itself a display. A 16x2 display module is shown from the front in [Figure 17-11](#) and from the rear in [Figure 17-12](#).

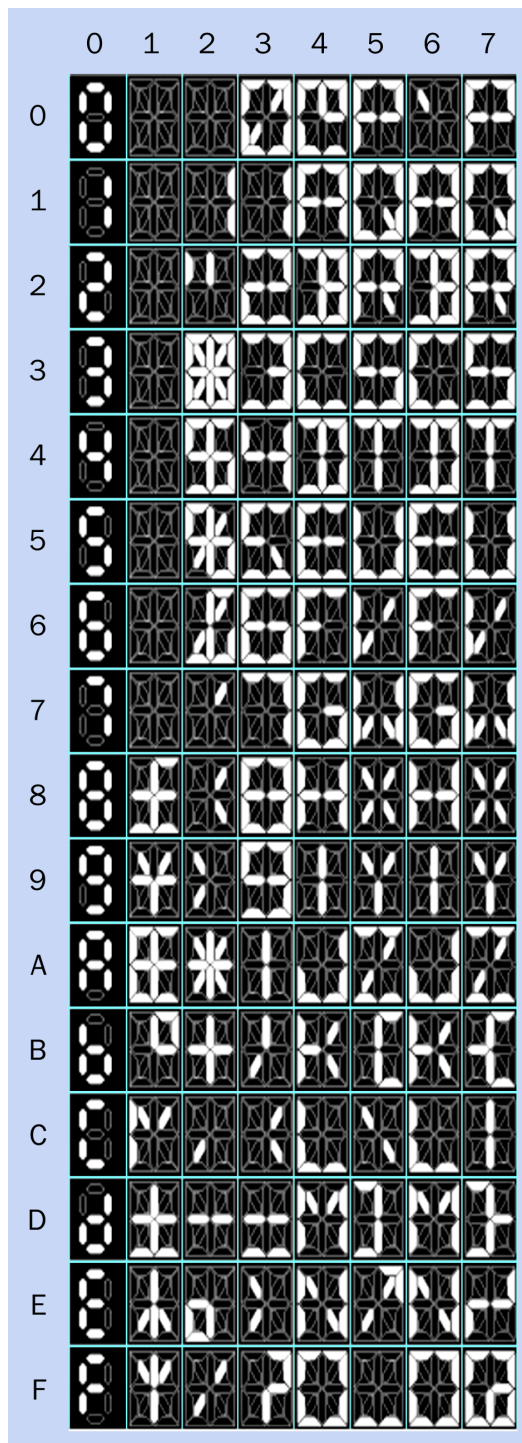


Figure 17-9. A full character set using 16-segment LCDs.

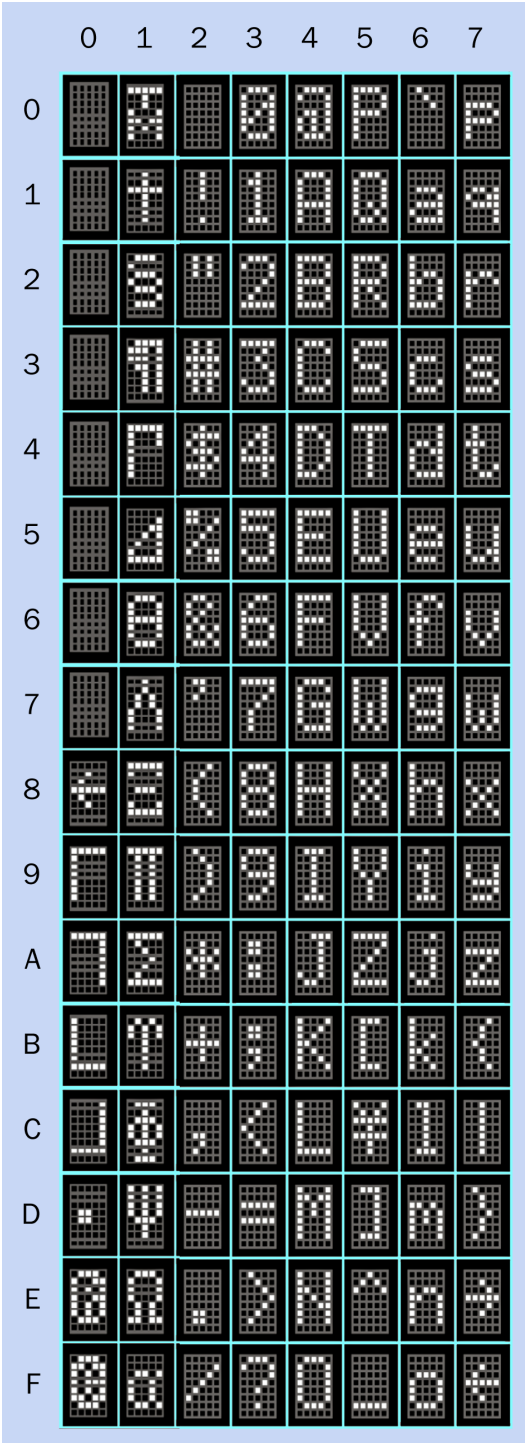


Figure 17-10. A dot-matrix character set typical of LCDs capable of displaying a matrix of 5×7 dots.

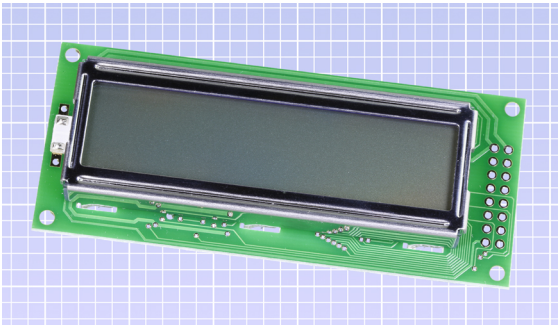


Figure 17-11. A 16x2 LCD display module seen from the front.

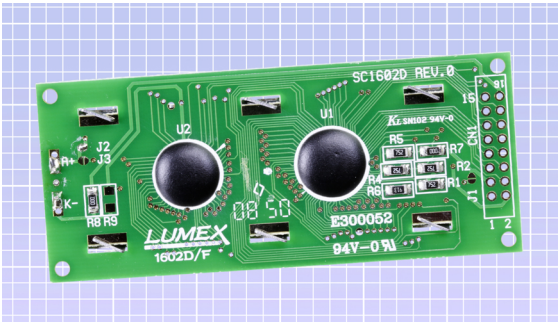


Figure 17-12. The same 16x2 LCD display module from the previous figure, seen from the rear.

Multiple-character display modules have been widely used in consumer electronics products such as audio components and automobiles where simple status messages and prompts are necessary—for example, to show the volume setting or broadcast frequency on a stereo receiver. Backlighting is almost always used.

Because the cost of small, full-color, high-resolution LCD screens has been driven down rapidly by the mass production of cellular phones, color displays are likely to displace monochrome dot-matrix LCD display modules in many applications. Similarly, touchscreens will tend to displace pushbuttons and tactile switches. Touchscreens are outside the scope of this encyclopedia.

Color

The addition of filters to create a full color display is shown in simplified form in [Figure 17-13](#).

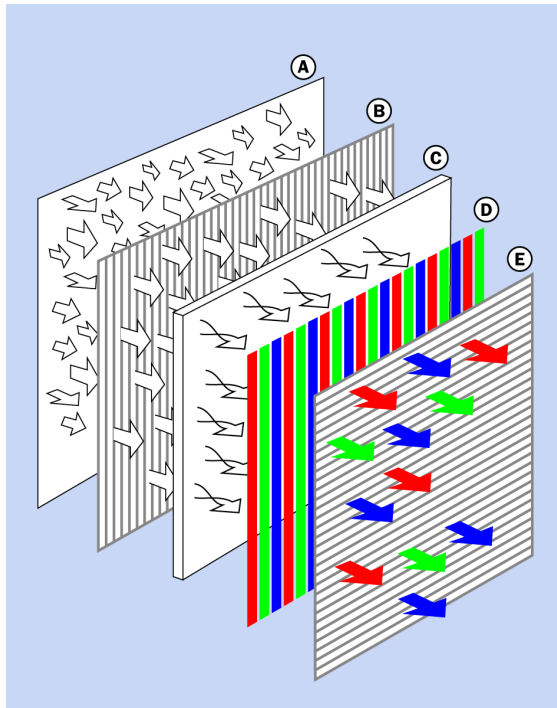


Figure 17-13. The addition of red, green, and blue color filters, in conjunction with variable density liquid crystal pixels, enables an LCD full-color display.

Red, green, and blue are almost always used as primary colors for transmitted light, because the combination of different intensities of these *RGB* primaries can create the appearance of many colors throughout the visible spectrum. They are said to be *additive* primaries, as they create brighter colors when they are combined. The principle is illustrated in [Figure 17-14](#).

The use of the word “primaries” to refer to red, green, and blue can cause confusion, as full-color printed materials use a different set of *reflective* primaries, typically cyan, magenta, and yellow, often with the addition of black. In this *CMYK* system, additional layers of pigment will absorb, or subtract, more visible frequencies. See [Figure 17-15](#).

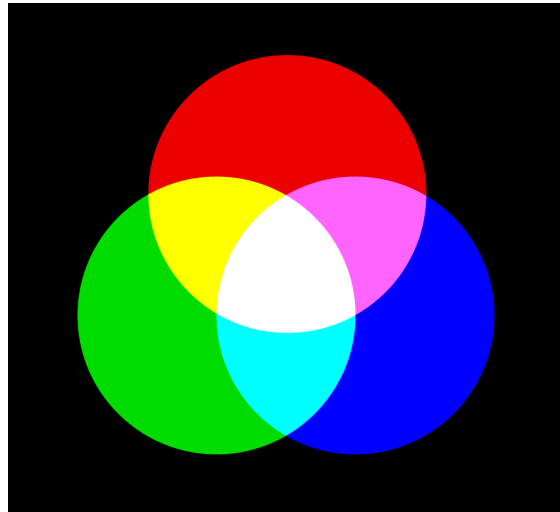


Figure 17-14. When colors red, green, and blue are transmitted directly to the eye, pairs of these additive primaries create secondary colors cyan, magenta, and yellow. Combining all three additive primaries creates an approximation of white light. This can be verified by viewing a color monitor with a magnifying glass.

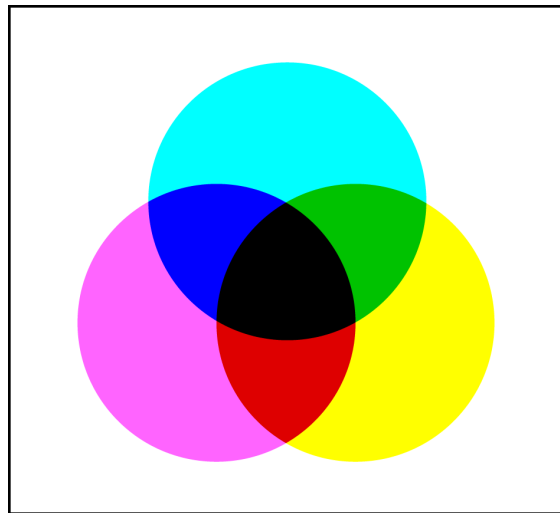


Figure 17-15. When ink colors cyan, magenta, and yellow are superimposed on white paper and are viewed in white light, pairs of these subtractive primaries create secondary colors red, green, and blue. Overprinting all three subtractive primaries creates an approximation of black, limited by the reflective properties of available pigments. Black ink is usually added to provide additional contrast.

The complete range of colors that can be created as a combination of primaries is known as the *gamut*. Many different RGB color standards have been developed, the two most widely used being *sRGB* (almost universal in web applications) and *Adobe 1998* (introduced by Adobe Systems for Photoshop, providing a wider gamut). None of the available systems for color reproduction comes close to creating the full gamut that can be perceived by the human eye.

Backlighting Options

For monochrome LCDs, **electroluminescent** backlighting may be used. It requires very low current, generates very little heat, and has a uniform output. However, its brightness is severely limited, and it requires an inverter that adds significantly to the current consumption.

For full-color LCDs, **fluorescent** lights were originally used. They have a long lifetime, generate little heat, and have low power consumption. However, they require a relatively high voltage, and do not work well at low temperatures. Early flat screens for laptop computers and desktop monitors used *cold-cathode fluorescent panels*.

Subsequently, white light-emitting diodes (**LEDs**) were refined to the point where they generated a range of frequencies that was considered acceptable. Light from the LEDs passes through a diffuser to provide reasonably consistent illumination across the entire screen. LEDs are cheaper than fluorescent panels, and allow a thinner screen.

High-end video monitors use individual red, green, and blue LEDs instead of a white backlight. This eliminates the need for colored filters and produces a wider gamut. So-called *RGB* LCD monitors are more expensive but are preferred for professional applications in video and print media where accurate color reproduction is essential.

Zero-Power Displays

Some techniques exist to create LCDs that require power only to flip them to and fro between

their transparent and opaque states. These are also known as *bistable* displays, but have not become as widely used. They are similar in concept to *e-ink* or *electronic paper* displays, but the principle of operation is different.

How to Use It

So long as an LCD consists of just one numeral, it can be driven by just one decoder chip that translates a binary-coded input into the outputs required to activate the appropriate segments of the LCD. The evolution of multi-digit displays, alphanumeric displays, dot-matrix displays, and graphical displays has complicated this situation.

Numeric Display Modules

An LCD consisting of a single digit is now a rare item, as few circuits require only one numeral for output. More commonly, two to eight numerals are mounted together in a small rectangular panel, three or four numerals being most common. A typical digital alarm clock uses a four-digit numeric display module, incorporating a colon and indicators showing AM/PM and alarm on/off. Other numeric display modules may include a minus sign.

Modules that are described as having 3.5 or 4.5 digits contain three full digits preceded by a numeral 1 composed of two segments. Thus, a 3-digit module can display numbers from 000 through 999, while a 3.5-digit display can display numbers from 000 through 1999, approximately doubling the range.

Numeric display modules of the type described here do not contain any decoder logic or drivers. An external device, such as a microcontroller, must contain a lookup table to translate a numeric value into outputs that will activate the appropriate segments in the numbers in a display, with or without decimal points and a minus sign. To avoid reinventing the wheel, a programmer may download code libraries for microcontrollers to drive commonly used numeric display modules. It is important to remember, though,

that segments in monochrome LCDs must be activated by AC, typically a square wave with a frequency of 30Hz to 90Hz.

An alternative is to use a decoder chip such as the 4543B or 4056B, which receives a binary-coded decimal input (i.e., 0000 through 1001 binary, on four input pins) and translates it into an output on seven pins suitable for connection with the seven segments of a 7-segment display. The 4543B requires a square-wave input to its “phase” pin. The square-wave must also be applied simultaneously to the backplane of the LCD, often identified as the “common” pin on datasheets. Pinouts for the 4543B are shown in [Figure 17-16](#).

The 4543B includes provision for “display blanking,” which can be used to suppress leading zeros in a multidigit number. However, the lack of outputs to control a minus sign or decimal point limits the decoder to displaying positive integers.

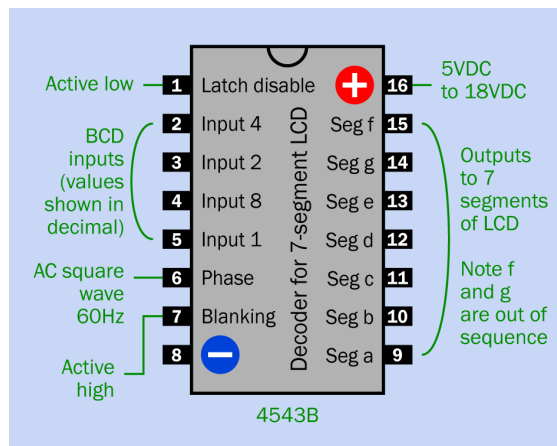


Figure 17-16. Pinouts for the 4543B decoder chip, which is designed to drive a seven-segment numeric LCD.

The power supply for a 4543B can range from 5VDC to 18VDC, but because the logic-high output voltage will be almost the same as that of the power supply, it must be chosen to match the power requirements of the LCD (very often 5VAC).(((

To drive a three-digit numeric display module, a separate decoder chip can be used to control each digit. The disadvantage of this system is that each decoder requires three inputs, so that a three-digit display will require nine outputs from the microcontroller.

To deal with this issue, it is common to [multiplex](#) a multi-digit display. This means that each output from the decoder is shared among the same segments of all the LCD numerals. Each LCD numeral is then activated in sequence by applying AC voltage to its common pin. Simultaneously, the decoder sends the data appropriate to that LCD. This process must be fast enough so that all the digits appear to be active simultaneously, and is best managed with a microcontroller. A simplified schematic is shown in [Figure 17-17](#). It can be compared with a similar circuit to drive LED displays, shown in [Figure 24-13](#).

Alphanumeric Display Module

Arrays of dot-matrix LCDs that can display alphabetical characters as well as numerals require preset character patterns (usually stored in ROM) and a command interpreter to process instructions that are embedded in the data stream. These capabilities are often built into the LCD module itself.

While there is no formal or de facto standard, the command set used by the Hitachi HD44780 controller is installed in many displays, and code libraries for this set are available for download from sites dedicated to the Arduino and other microcontrollers. Writing code from scratch to control all aspects of an alphanumeric display is not a trivial chore. The Hamtronix HDM08216L-3-L30S is a display that incorporates the HD44780.

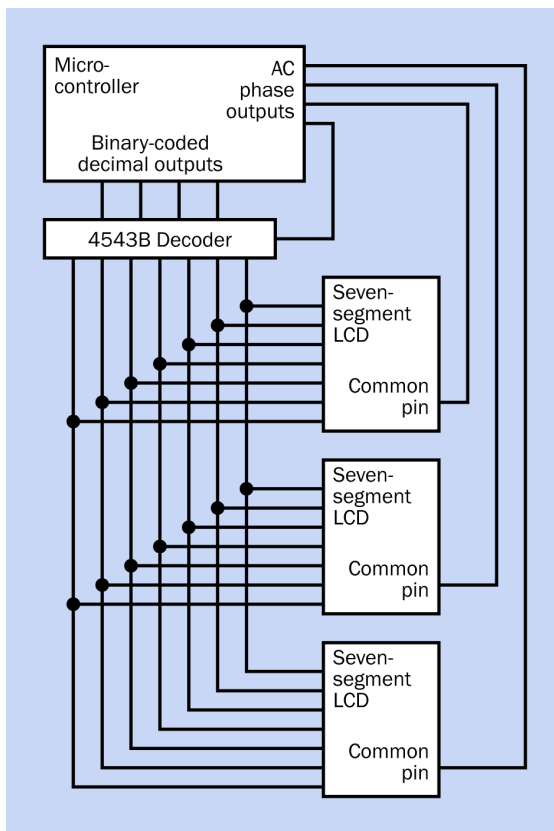


Figure 17-17. When two or more numeric displays are multiplexed, a control device (typically, a microcontroller) activates each of them in turn via its backplane (common terminal) while sending appropriate data over a shared bus.

Regardless of which standard is used, some features of alphanumeric display modules are almost universal:

- Register select pin. Tells the display whether the incoming data is an instruction, or a code identifying a displayable character.
- Read/write pin. Tells the display whether to receive characters from a microcontroller or send them to a microcontroller.
- Enable/disable pin.
- Character data input pins. There will be eight pins to receive the 8-bit ASCII code for each displayable character in parallel. Often there is an option to use only four of these pins, to

reduce the number of microcontroller outputs necessary to drive the display. Where four pins are used, each 8-bit character is sent in two segments.

- LED backlight pin. Two may be provided, one connected to the anode(s) of the LED backlight, the other to the cathode(s).
- Reset pin.

Embedded instruction codes can be complex, including commands to reposition the cursor at a specific screen location, backspace-and-erase, scroll the display, and erase all characters on the screen. Codes may be included to adjust screen brightness and to switch the display between light-on-dark (negative) and dark-on-light (positive) characters.

Some display modules also have graphics capability, allowing the user to address any individual pixel on the screen.

Because of the lack of standardization in control codes, manufacturer's datasheets must be consulted to learn the usage of a particular alphanumeric display module. In addition to datasheets, online user forums are a valuable source of information regarding quirks and undocumented features.

What Can Go Wrong

Temperature Sensitivity

Liquid crystals vary in their tolerance for low and high temperatures, but generally speaking, a higher voltage may be necessary to create a sufficiently dense image at a low temperature. Conversely, a lower voltage may be necessary to avoid "ghosting" at a high temperature. An absolutely safe operating temperature range is likely to be 0 through 50 degrees Celsius, but check the manufacturer's datasheet for confirmation. Special-purpose LCDs are available for extreme temperatures.

Excessive Multiplexing

A twisted nematic display is likely to perform poorly if its duty cycle is greater than 1:4. In other words, more than four displays should not be multiplexed by the same controller.

DC Damage

An LCD can be damaged quickly and permanently if it is subjected to DC current. This can occur by accident if, for example, a timer chip is being used to generate the AC pulse stream, and the timer is accidentally disconnected, or has an incorrect connection in its RC network. Check timer output with a meter set to measure AC volts before allowing any connection to the common pin of an LCD.

Bad Communications Protocol

Many alphanumeric display modules do not use a formal communications protocol. Duplex serial

or I2C connection may not be available. Care must be taken to allow pauses of a few milliseconds after execution of embedded commands, to give the display sufficient time to complete the instruction. This is especially likely where a command to clear all characters from the screen has to be executed. If garbage characters appear on the screen, incorrect data transfer speed or lack of pause times may be to blame.

Wiring Errors

This is often cited by manufacturers as the most common cause of failure to display characters correctly, or lack of any screen image at all.

incandescent lamp

The terms *incandescent light*, *incandescent bulb*, and *incandescent light bulb* are often used interchangeably with **incandescent lamp**. Because the term “lamp” seems to be most common, it is used here. A *panel-mounted indicator lamp* is considered to be an assembly containing an incandescent lamp.

A *carbon arc*, which generates light as a self-sustaining spark between two carbon electrodes, can be thought of as a form of incandescent lamp, but is now rare and is not included in this encyclopedia.

OTHER RELATED COMPONENTS

- **LED area lighting** (see [Chapter 23](#))
- **LED indicator** (see [Chapter 22](#))
- **neon bulb** (see [Chapter 19](#))
- **fluorescent light** (see [Chapter 20](#))

What It Does

The term *incandescent* describes an object that emits visible light purely as a consequence of being hot. This principle is used in an **incandescent lamp** where a wire *filament* glows as a result of electric current passing through it and raising it to a high temperature. To prevent oxidation of the filament, it is contained within a sealed bulb or tube containing an inert gas under low pressure or (less often) a vacuum.

Because incandescent lamps are relatively inefficient, they are not considered a wise environmental choice for area lighting and have been prohibited for that purpose in some areas. However, small, low-voltage, panel-mount versions are still widely available. For a summary of advantages of miniature incandescent lamps relative to **light-emitting diodes** (LEDs) see “[Relative Advantages](#)” on page 179.

Schematic symbols representing an incandescent lamp are shown in [Figure 18-1](#). The symbols

are all functionally identical except that the one at bottom right is more likely to be used to represent small panel-mounted indicators.

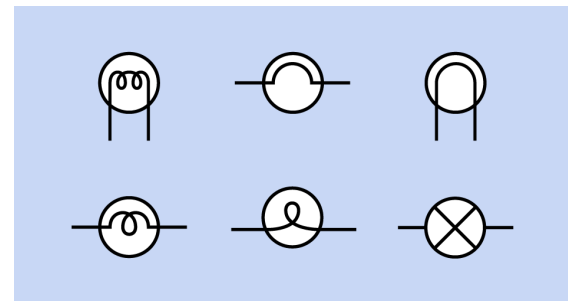


Figure 18-1. A variety of symbols can represent an incandescent lamp. The one at bottom right may be more commonly used for small panel-mounted indicators.

The parts of a generic incandescent light bulb are identified in [Figure 18-2](#):

A: Glass bulb.

B: Inert gas at low pressure.

C: Tungsten filament.

D: Contact wires (connecting internally with brass base and center contact, below).

E: Wires to support the filament.

F: Internal glass stem.

G: Brass base or cap.

H: Vitreous insulation.

I: Center contact.

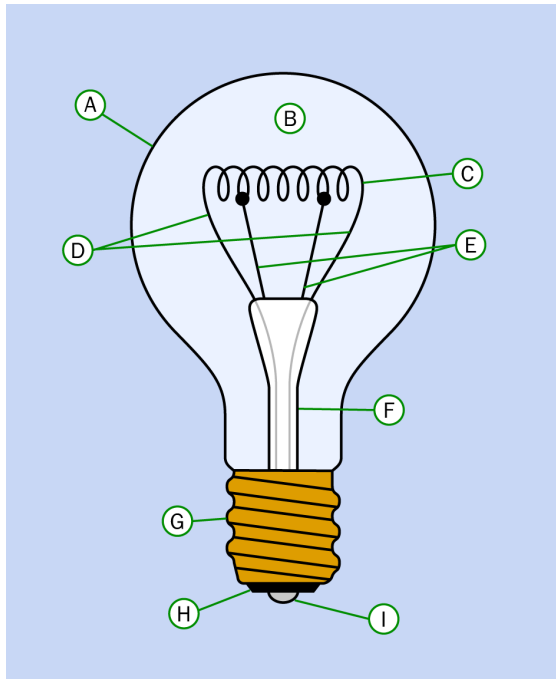


Figure 18-2. The parts of a typical incandescent lamp (see text for details).

History

The concept of generating light by using electricity to heat a metal originated with Englishman Humphrey Davy, who demonstrated it with a large battery and a strip of platinum in 1802. Platinum was thought to be suitable because it has a relatively high melting point. The lamp worked but was not practical, being insufficiently bright and having a short lifespan. In addition, the platinum was prohibitively expensive.

The first patent for an incandescent lamp was issued in England in 1841, but it still used platinum. Subsequently, British physicist and chemist Joseph Swan spent many years attempting to develop practical carbon filaments, and obtained a patent in 1880 for *parchmentized thread*. His house was the first in the world to be illuminated by light bulbs.

Thomas Edison began work to refine the electric lamp in 1878, and achieved a successful test with a carbonized filament in October 1879. The bulb lasted slightly more than 13 hours. Lawsuits over patent rights ensued. Carbonized filaments were used until a tungsten filament was patented in 1904 by the German/Hungarian inventor Just Sándor Frigyes and the Croatian inventor Franjo Hanaman. This type of bulb was filled with an inert gas, instead of using a vacuum.

Many other pioneers participated in the effort to develop electric light on a practical basis. Thus it is incorrect to state that “Thomas Edison invented the light bulb.” The device went through a very lengthy process of gradual refinement, and one of Edison’s most significant achievements was the development of a power distribution system that could run multiple lamps in parallel, using filaments that had a relatively high resistance. His error was insisting on using *direct current* (DC) while his rival Westinghouse pioneered *alternating current* (AC), enabling power transmission over longer distances through the use of **transformers**. The use of AC also enabled Tesla’s brushless induction motor.

By the mid-1900s, most incandescent bulbs used tungsten filaments.

How It Works

All objects emit electromagnetic radiation as a function of their temperature. This is known as *black body radiation*, based on the concept of an object that absorbs all incoming light, and thus does not reflect any sources from outside itself. As its temperature increases, the intensity of the

radiation increases while the wavelength of the radiation tends to decrease.

If the temperature is high enough, the wavelength of the radiation enters the visible spectrum, between 380 and 740 nanometers. (A nanometer is one-billionth of a meter.)

The melting point of tungsten is 3,442 degrees Celsius, but a lamp filament typically operates between 2,000 and 3,000 degrees. At the higher end of this scale, evaporation of metal from the filament tends to cause deposition of a dark residue on the inside of the bulb, and erodes the filament more rapidly, to the point where it eventually breaks. At the lower end of this scale, the light will be yellow and the intensity will be reduced.

Spectrum

The color of black-body radiation is measured using the Kelvin temperature scale. The increment of 1 degree Kelvin is the same as 1 degree Celsius, but the Kelvin scale has a zero value at *absolute zero*. This is the theoretical lowest conceivable temperature, at which there is complete absence of heat. It is approximately -273 degrees Celsius.

From this it is evident that if K is a temperature in degrees Kelvin and C is a temperature in degrees Celsius:

$$K = C + 273 \text{ (approximately)}$$

Calibration of light sources in degrees Kelvin is common in photography. Many digital cameras allow the user to specify the *color temperature* of lights that are illuminating an indoor scene, and the camera will compensate so that the light source appears to be pure white with all colors in the visible spectrum being represented equally.

Some computer monitors also allow the user to specify a white value in degrees Kelvin.

Color temperature is used in astronomy, because the spectrum of many stars is comparable with that of a theoretical black body.

A color temperature of 1,000 degrees K will have a dark orange hue, while 15,000 degrees K or higher will have a blue hue comparable to that of a pale blue sky. The color temperature of the sun is approximately 5,800 K. Interior lighting is often around 3,000 K, which many people find acceptable because it creates pleasant flesh tones. An incandescent bulb described by the manufacturer as “soft white” or “warm” will have a lower color temperature than one which is sold as “pure white” or “paper white.”

Graphs showing the emission of wavelengths at various color temperatures are shown in [Figure 18-3](#). The rainbow section indicates the approximate range of visible wavelengths between ultraviolet, on the left, and infrared, on the right. For purposes of clarity, the peak intensity for each color temperature has been equalized. In reality, increasing the temperature also increases the light output.

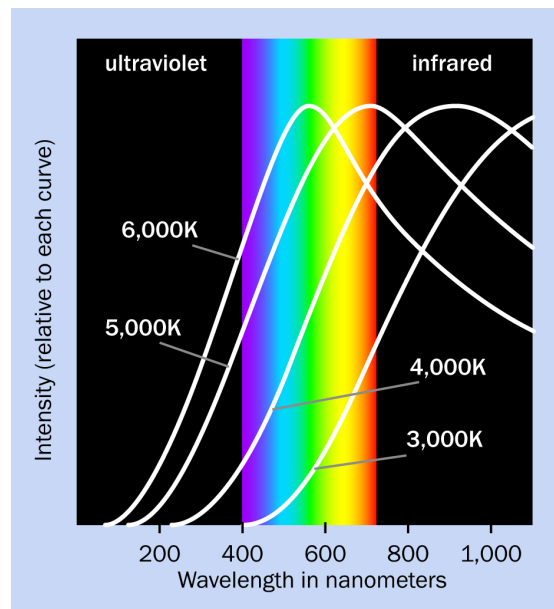


Figure 18-3. Approximate peak wavelengths for black-body radiation at various color temperatures in degrees Kelvin. The curves have been adjusted so that their peak values are equalized. Adapted from an illustration in the reference book *Light Emitting Diodes* by E. Fred Schubert.

Non-Incandescent Sources

So long as light is generated by heating a filament, plotting the intensity against wavelength will result in a smooth curve without irregularities. A higher Kelvin value will simply displace and compress the curve laterally without changing its basic shape to a significant degree.

The introduction of **fluorescent** sources and, subsequently, light-emitting diodes (**LEDs**) has complicated this scenario. Because they are *luminescent* rather than *incandescent*, they do not generate an evenly weighted, continuous range of wavelengths.

LEDs tend to emit monochromatic light, meaning that it is tightly centered around just one color. A “white” LED is really a blue LED in which a phosphor coating on the semiconductor die is excited to create light over a broader range. A fluorescent light tends to create *spectral lines* which show up as sharp peaks at a few wavelengths determined by the mercury inside the bulb. [Figure 18-4](#) illustrates these problems.

The human eye tends to compensate for the yellow emphasis of incandescent lamps and for the irregularities in spectra emitted by other light sources. Also, the eye is often unable to distinguish between “white” light created as a mix of all the visible wavelengths, and light that appears white even though it is dominated by a few isolated wavelengths from a fluorescent source.

However, when the eye views colors that are illuminated by a source that has gaps in its spectrum, some of the colors will appear unnaturally dull or dark. This is true also if an imperfect source is used as a backlight to create colors on a video monitor. Colors rendered by different light sources are shown in [Figure 23-7](#) and subsequent figures.

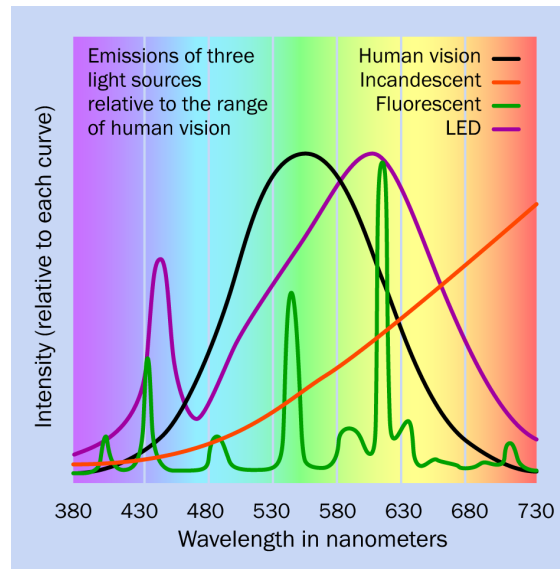


Figure 18-4. The relative performance of three light sources compared with sensitivity of the human eye to the visible spectrum. Note that the range of wavelengths on the horizontal scale in this figure is not the same as the range in the previous figure. The color assigned to each curve is arbitrary. Adapted from VU1 Corporation.

Photography is adversely affected by the use of LEDs or fluorescents as a light source. Reds, for example, can seem dark when lit by white LEDs, while blues can be inappropriately intense. Because the source does not have an emission curve comparable to that of an incandescent light, the auto-white balance feature of a digital camera may be unable to address this problem, and it cannot be resolved by entering a different Kelvin number manually.

The fidelity with which a light source is capable of displaying the full visible spectrum is known as the *color rendering index* (CRI), ranging from a perfect score of 100 down to 0 or even lower (sodium-vapor street lighting has a negative value). Computing the index requires standard reference color samples and has been criticized for generating scores that do not correlate well with subjective assessments.

Incandescent bulbs can have a CRI of 100, while an uncorrected “white” LED may score as low as 80.

Power Consumption

Approximately 95% of the power consumed by an incandescent lamp generates heat instead of visible light. This wastage of power in room lighting is compounded by the power consumption of air conditioning to remove the heat from enclosed spaces in hot climates. While the heat from incandescent lamps does reduce the need for space heating in cold environments, heat is delivered more efficiently by using systems designed for that purpose. Consequently, greater energy efficiency can be achieved with a light source that generates less heat, regardless of ambient air temperature.

Variants

Miniature Lamps

Prior to the development of **LEDs**, all light-emitting panel-mounted indicators were either **neon bulbs** or **incandescent lamps**. The use of neon is limited by its need for a relatively high voltage.

Miniature incandescents were the traditional choice for battery-powered light sources, and at the time of writing are still used in cheap flashlights. Variants are available that are as small as a 5mm LED, with a claimed life expectancy that is comparable, although they draw more current to generate an equivalent light intensity, because much of their power is wasted in infrared wavelengths.

The photograph in [Figure 18-5](#) is of a miniature lamp terminating in pins spaced 0.05" apart. The total height of the lamp, including its ceramic base, is less than 0.4," while its diameter is just over 0.1". It draws 60mA at 5V and is rated for 25,000 hours.

The photograph in [Figure 18-6](#) is of a lamp of similar size and power consumption, but terminating in wire leads and rated for 100,000 hours. It emits 0.63 lumens.

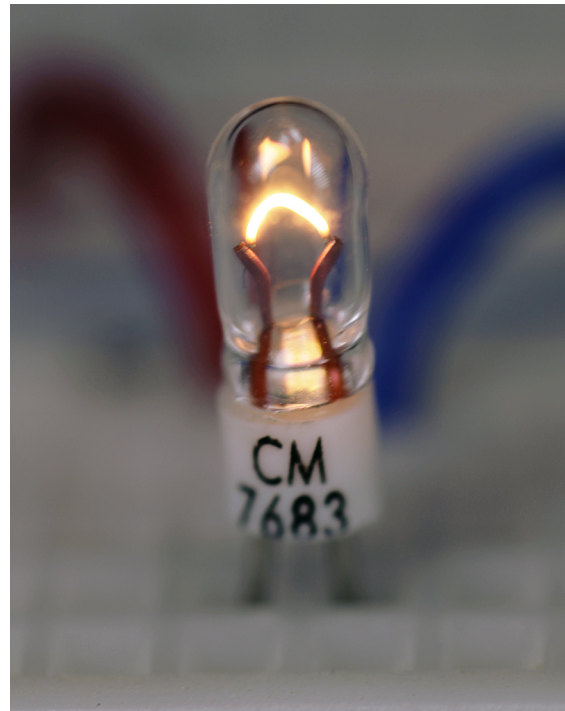


Figure 18-5. A miniature lamp less than 0.4" high, terminating in pins spaced 0.05" apart.



Figure 18-6. This lamp is 0.25" high and terminates in wire leads.

The lamp pictured in [Figure 18-7](#) is slightly larger, with a glass envelope about 0.25" diameter. It is rated for less than half the lifetime of the lamp in [Figure 18-6](#) but emits three times as much light—a typical tradeoff. Various base styles are available.



Figure 18-7. This lamp has a glass envelope about 0.35" high. Its screw-in base makes it easier to replace than an LED.

In the United States, the light output from miniature incandescent lamps may be measured in lumens, but is more often rated in [mean spherical candlepower](#) (MSCP). An explanation of light measurement is included in ["MSCP" on page 178](#).

[Lamp lenses](#) provide a quick and simple way to add color to a miniature incandescent lamp. Usually the lens is cylindrical with a hemispherical end cap, and is designed to push-fit or snap-fit over a small lamp. Even when the cap is translucent, it may still be referred to as a lens.

Panel-Mount Indicator Lamps

This term often refers to a tubular assembly containing a miniature lamp, ready for installation.

The enclosure is often designed to snap-fit into a hole drilled in the panel. If the incandescent bulb inside the enclosure cannot be replaced, the component is said to be "non-relampable." [Figure 18-8](#) shows a 12-volt panel-mount indicator lamp.

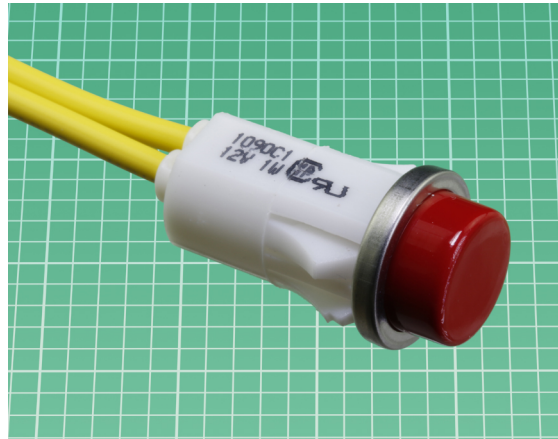


Figure 18-8. This panel-mount indicator lamp is designed to push-fit into a hole 1/2" in diameter. The bulb inside it is not replaceable, causing the assembly to be classified as "non-relampable."

Halogen or Quartz-Halogen

This is a type of incandescent lamp containing gases under pressure in which halogens such as iodine or bromine cause evaporated tungsten atoms to be redeposited on the filament. A halogen lamp can therefore operate at a higher temperature, creating a light that is less yellow and brighter than that from a comparable incandescent lamp. It also enables a smaller bulb, but requires an envelope of borosilicate-halide glass (often termed [fused quartz](#)) instead of regular glass. A halogen lamp will be slightly more efficient than an incandescent bulb of the same wattage, and will last longer.

Halogens are available in a variety of formats. The small bulb pictured in [Figure 18-9](#) consumes 75W, emitting 1,500 lumens at 3,000 degrees Kelvin. The light intensity is claimed to be equivalent to that of a 100W incandescent bulb. It has a mini-candelabra base.

Oven Lamps

Oven lamps are designed to withstand the high temperature in an oven. Typically they are usable with ambient temperatures up to 300 degrees C. A common power rating is 15W.



Figure 18-9. A halogen bulb slightly more than 2" in length, designed for 115VAC.

Base Variants

Miniature lamps are available with a wide variety of connection options, including wire terminals, single-contact bayonet, double-contact bayonet, miniature screw base, and fuse style. Most of these options require a matching socket.

Screw-in lamps for room illumination are common in household lighting in the United States and many other countries (but not in the UK, where bayonet fittings are used). The US socket size is designated by letter E followed by a number that gives the socket diameter in millimeters. Common sizes are E10, E14 and E27.

A *bayonet base* is fitted with two small lugs protruding on opposite sides. The lamp is secured by pushing it in and twisting it to engage the lugs in slots in the socket. The advantage of a bayonet

base is that the bulb is less likely to become loose as a result of vibration.

A *pin base* consists simply of a pair of pins that will push-fit into small holes in a socket.

A *flange base* has a flange that engages in a socket where flexible segments will retain it.

A *wedge base* is forced between two contacts which retain the bulb by friction.

Some indicator lamps terminate simply in long, thin leads that can be soldered.

Values

While the power consumption of full-size incandescent lamps is rated in watts, small indicator lamps are rated in milliamperes at the voltage for which they are designed. Miniature lamps may require specific voltages ranging from as low as 2V to 24V. A higher voltage generally necessitates a longer filament, which may entail a larger bulb.

The light that a lamp will emit can be measured in two ways: either as the power of the lamp (not its power consumption, but its radiating power), or as the light delivered to a specific area at a specific distance. These two measurements may differ because a lamp may concentrate its light in a beam, as in the case of a reflector bulb or an LED.

Power

Flux, in watts, is a measurement of energy flow in joules per second. The total radiating power of a lamp, in all wavelengths, in all directions, is known as its *radiant flux*. Because invisible wavelengths are of little interest when assessing the brightness of a lamp, the term *luminous flux* is used to describe the apparent brightness of the lamp in the visible spectrum. The unit for luminous flux is the *lumen*.

The human eye is most responsive to yellow-green hues in the center of the spectrum. Consequently, the measurement of luminous flux is

weighted toward green at a wavelength of 555 nanometers. Red and violet are considered to have low luminous flux, while infrared and ultra-violet have a zero value.

When considering a value expressed in lumens, remember:

- Lumens are a measure of the total radiated power output of a light source, in all directions, in the visible spectrum only, weighted toward the characteristics of the human eye.
- The number of lumens of a light source does not define the direction in which the light is shining, or its uniformity.
- The abbreviation for lumen is *lm*.

A conventional incandescent lamp that consumes 100W of electricity is likely to have a light output of about 1,500 lumens. A 40W fluorescent tube can have a light output of about 2,600 lumens.

Illuminance

The *illuminance* of a light source is defined as the luminous flux per unit of area. This can be thought of as the brightness of a surface illuminated by the source.

Illuminance is measured in *lux*, where 1 lux = 1 lumen per square meter. For accurate calibration, the illuminated surface should be spherical in shape, and must be located 1 meter from the light source, with the source at the geometrical center of the sphere.

Illuminance used to be measured in *foot-candles*, where 1 foot-candle was 1 lumen per square foot.

- The number of lumens per square meter (lux) does not define the size of the illuminated area, only the brightness per unit of area.
- A lamp that has a tightly focused beam can achieve a high lux rating. When selecting a lamp for an application, the angle of disper-

sion of the beam must be considered in conjunction with its lux rating.

Intensity

A *candela* measures the luminous flux within an angle of dispersion. The angle is three-dimensional, and can be imagined as the sharpness of a point of a cone, where the light source is at the point and the cone represents the dispersion of light.

The three-dimensional angle of dispersion is measured in *steradians*. If a light source is at the center of a sphere that has a radius of 1 meter, and is illuminating one square meter of the surface of the sphere, the angle of dispersion is 1 steradian.

- A source of 1 lumen which projects all its light through a dispersion angle of 1 steradian is rated at 1 candela.
- The number of candelas does not define the angle of dispersion, only the intensity within that angle.
- A light source rated for 1,000 candelas could have a power of 10 lumens concentrated within an angle of 0.01 steradians, or could have a power of only 1 lumen concentrated within an angle of 0.001 steradians.
- There are 1,000 millicandelas in 1 candela. The abbreviation for candela is *cd* while the abbreviation for a millicandela is *mcd*.
- LEDs are often rated in mcd. The number describes the intensity of light within its angle of dispersion.

MSCP

Although the term *candlepower* is obsolete, it has been redefined as being equal to 1 candela. *Mean spherical candlepower* (MSCP) is a measurement of all the light emitted from a lamp in all directions. Because the light is assumed to be omnidirectional, it fills $4 * \pi$ (about 12.57) steradians. Therefore 1 MSCP = approximately 12.57

lumens. In the United States, MSCP is still the most common method of rating the total light output of a miniature lamp.

Efficacy

The *radiant luminous efficacy* (abbreviated *LER*) assesses how effective a lamp is at channeling its output within the visible spectrum, instead of wasting it in other wavelengths, especially infrared. LER is calculated by dividing the power emitted in the visible spectrum (the *luminous flux*) by the power emitted over all wavelengths.

Thus, if VP is the power emitted in the visible spectrum, and AP is the power emitted in all wavelengths:

$$\text{LER} = \text{VP} / \text{AP}$$

LER is expressed in *lumens per watt*. It can range from a low value of around 12 lm/W for a 40W incandescent bulb to 24 lm/W for a quartz halogen lamp. Fluorescent lamps may average 50 lm/W. LEDs vary, but can achieve 100 lm/W.

Efficiency

The *radiant luminous efficiency* (abbreviated *LFR*) of a lamp measures how good its radiant luminous efficacy is, compared with an imaginary ideal lamp. (Note the difference between the words “efficiency” and “efficacy.”) LFR is determined by dividing the radiant luminous efficacy (LER) by the maximum theoretical LER value of 683 lm/W, and multiplying by 100 to express the result as a percentage. Thus:

$$\text{LFR} = 100 * (\text{LER} / 683)$$

The LFR ranges from around 2% for a 40W bulb to 3.5% for a quartz halogen lamp. LEDs may be around 15% while fluorescents are closer to 10%.

How to Use It

When first introduced, LEDs were limited by their higher price, lower maximum light output, and inability to display blue or white. The price difference has disappeared for small indicators, while gaps in the color range have been filled

(although the color rendering index of LEDs is still inferior).

Brightness remains an advantage for large incandescents relative to LEDs, as they are more upwardly scalable. However, fluorescents and vapor lamps have an advantage for very high light output, as in the lighting of big-box stores or parking lots. Thus the range of applications for incandescent bulbs is diminishing, especially because common types are now illegal for domestic light fixtures in many parts of the world.

Relative Advantages

When choosing whether to use an incandescent lamp or an LED, these advantages of an incandescent lamp should be considered:

- The intensity can be adjusted with a **triac**-based dimmer. Regular fluorescents cannot be dimmed, while LEDs often require different dimmer circuitry.
- The intensity can also be adjusted with a rheostat. The output from fluorescents cannot.
- Easy white-balance correction. LEDs and fluorescents do not naturally produce a consistent output over the visible spectrum.
- Can be designed to operate directly from a wide range of voltages (down to around 2V and up to around 300V). A higher voltage entails a longer filament wire, which may require a larger bulb. LEDs require additional components and circuitry to use higher voltages.
- Incandescent bulbs are more tolerant of voltage fluctuations than LEDs. With battery operation, the incandescent will still provide some reduced light output when the voltage has diminished radically. LEDs will not perform at all at currents lower than their threshold.
- An incandescent is nonpolarized and may be socketed, which simplifies user replacement.

LEDs are polarized and are usually soldered in.

- Can be powered by AC or DC without any modification or additional circuitry. LEDs require DC, which must be provided through a transformer and rectifier, or similar electronics, if AC power is the primary source.
- Can be equally visible from a wide range of viewing angles. LEDs have restricted viewing angles.
- The heat output from an incandescent bulb may occasionally be useful (for example in a terrarium, or in incubators for poultry).
- Trouble-free switching. Fluorescents tend to hesitate and blink when power is applied, and they require a *ballast* to energize them. The lifespan of fluorescents is reduced by frequent switching.
- No low-temperature problems. Incandescent lamps are not significantly affected by low temperatures. Fluorescents may not start easily in a cold environment, and may flicker or glow dimly for 10 minutes (or more) until they are warm enough to function properly.
- Easy disposal. Fluorescent lights contain small quantities of mercury that are an environmental hazard. They should not be mixed with ordinary trash. Compact fluorescent lamps (CFLs) and LEDs used for room lighting will be packaged with electronics that should ideally be recycled, although this is not very practical. Incandescent bulbs impose the least burden on the environment when they are thrown away.

However, the incandescent lamp has some obvious disadvantages:

- Relatively inefficient.
- More susceptible to vibration.
- More fragile.
- Likely to have a shorter natural life expectancy than LEDs, fluorescents, or neon bulbs,

although the lifetime of a small panel indicator can be equal to that of an LED if a low color temperature is acceptable.

- Requires a filter or tinted glass envelope to generate colored light. This further reduces the lamp's efficiency.
- Cannot be miniaturized to the same degree as an LED indicator.

Derating

The lifespan of a lamp can be greatly extended by choosing one with a higher current rating or using it at a lower voltage. The light output will be reduced, and the color temperature will be at a lower Kelvin number, but in some situations this tradeoff may be acceptable.

The graphs in [Figure 18-10](#) suggest that if the voltage of a hypothetical miniature lamp is reduced to 80% of the manufacturer's recommended value, this can make the lamp last 20 times as long. Note, however, that this will cut the light intensity to 50% of its normal value.

Conversely, using 130% normal voltage will give 250% of the normal light output, while shortening the life of the lamp to 1/20 of its normal value. Naturally these figures are approximations that may not apply precisely to a specific lamp.

What Can Go Wrong

High Temperature Environment

If an incandescent lamp is used in an environment hotter than 100 degrees Celsius, the life of the lamp is likely to be reduced by the "water cycle." Any water molecules inside the glass envelope will break down, allowing oxygen to combine with the tungsten filament to form tungsten oxide. The tungsten is deposited on the inside of the glass while the oxygen is liberated and begins a new cycle.

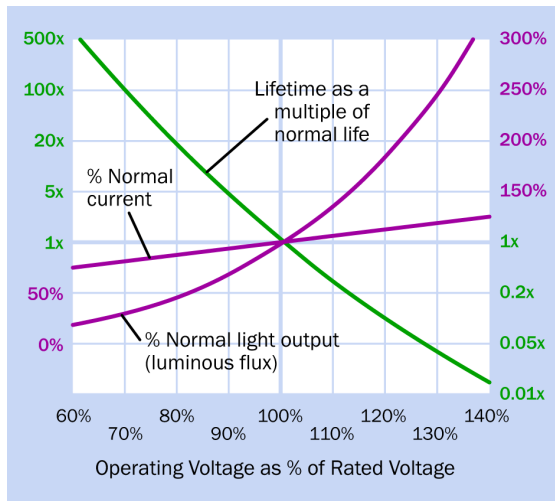


Figure 18-10. The life expectancy of a hypothetical miniature lamp is very strongly influenced by voltage. Applying only 60% of the rated voltage can make a lamp last 500 times its normal lifespan, although it will greatly reduce light output. Note that the vertical axes apply to curves of the same color. Adapted from “Characteristics of Miniature Lamps” from Toshiba Lighting and Technology Corporation.

Fire Risk

The partially evacuated bulb of an incandescent lamp provides some separation and protection from the heat in the filament, but if the bulb cannot disperse heat by radiation or convection, its temperature can rise to the point where it ignites flammable materials.

Halogen lamps have an elevated fire risk because they operate at a higher temperature and are

smaller, providing less surface area to disperse the heat. They also contain gases under seven to eight atmospheres of pressure. Thermal stress can cause a halogen bulb to shatter, and fingerprints on the glass can increase this risk.

Current Inrush

When an incandescent lamp is first switched on, its filament has one-tenth the resistance that it will exhibit when it becomes hot. Consequently, the lamp will take a large initial surge of current, which stabilizes after about 50 milliseconds. This should be considered if one or more small lamps shares a DC power supply with components such as logic chips that may be sensitive to voltage fluctuations.

Replacement Problems

Because of the limited life of incandescent lamps, they should be installed in such a way that they are easy to replace. This can be an issue with panel indicators, where disassembly of a device may be necessary to reach the lamp.

The range of small incandescent lamps is diminishing, and may continue to diminish in the future. Future availability of replacement lamps should be considered when designing a circuit. When building equipment in small quantities, spare lamps should be purchased for future use.

neon bulb

The terms **neon bulb**, *neon indicator*, and *neon lamp* tend to be used interchangeably. In this encyclopedia, a *neon bulb* is defined as a glass capsule containing two electrodes in neon gas (or a combination of gases in which neon is present). A *neon lamp* is an assembly containing a neon bulb, usually using a plastic tube with a tinted transparent cap at one end. A *neon indicator* is a miniature neon lamp that is usually panel-mounted.

Large-scale neon tubes used in signage are not included in this encyclopedia.

OTHER RELATED COMPONENTS

- **incandescent lamp** (see [Chapter 18](#))
- **fluorescent light** (see [Chapter 20](#))
- **LED indicator** (see [Chapter 22](#))

What It Does

When voltage is applied between two electrodes inside a neon bulb, the inert gas inside the bulb emits a soft red or orange glow. This color may be modified by using a tinted transparent plastic cap, known as a *lens*, in a *neon lamp* assembly.

A neon bulb is usually designed for a power supply of 110V or higher. It functions equally well with alternating or direct current.

The schematic symbols in [Figure 19-1](#) are commonly used to represent either a neon bulb or a neon lamp. They are all functionally identical. The black dot that appears inside two of the symbols indicates that the component is gas filled. The position of the dot inside the circle is arbitrary. Even though all neon bulbs are gas filled, the dot is often omitted.

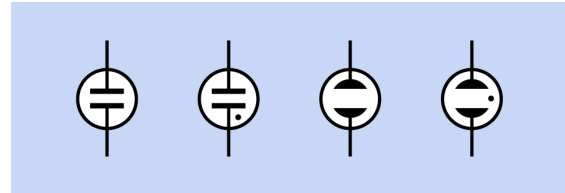


Figure 19-1. Any of these symbols may represent a neon bulb or a neon lamp. The dot in two of the symbols indicates that the component is gas filled. All neon bulbs are gas filled, but the dot is often omitted.

The photograph in [Figure 19-2](#) shows a neon bulb with a series resistor preattached to one lead. Many bulbs are sold in this configuration, because a resistor must be used to limit current through the bulb. The bulb has no polarity and can be used on an AC or DC power supply. The same bulb is shown in its energized state in [Figure 19-3](#).

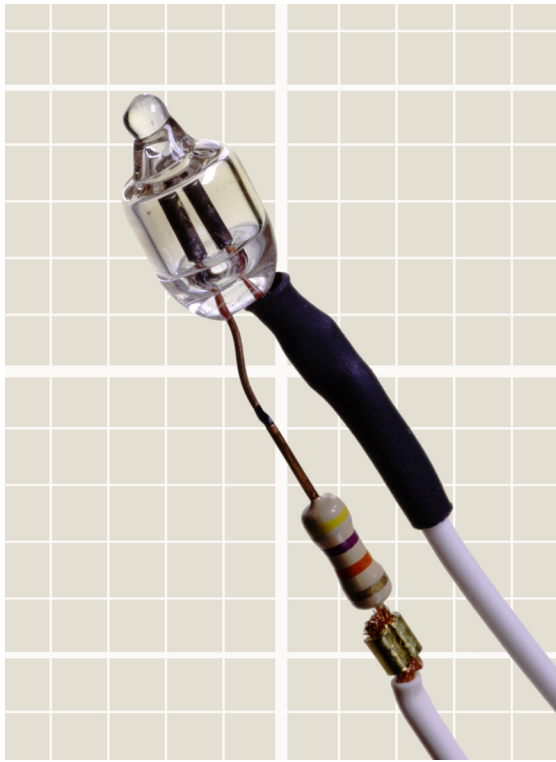


Figure 19-2. A typical neon bulb with series resistor attached to one lead.

How It Works

Construction

The parts of a neon bulb are illustrated in [Figure 19-4](#). When the bulb is fabricated, it begins as a glass tube. The leads are made of *dumet*, consisting of a copper sheath around a nickel iron core. This has the same coefficient of expansion as glass, so that when the glass is heated and melted around the leads, it forms a seal that should be unaffected by subsequent temperature fluctuations. This area is known as the *pinch* in the tube.

Nickel electrodes are welded onto the leads before the leads are inserted into the tube. The electrodes have an emissive coating that reduces the minimum operating voltage. The glass tube is filled with a combination of neon and argon gases, or pure neon for higher light output (which reduces the life of the component). The

top end of the glass tube is heated until it melts, and is pinched off. This creates a distinctive protrusion known as the *pip*.



Figure 19-3. The same bulb from the previous photograph, energized with 115VAC.

Ionization

When a voltage is applied between the leads to the bulb, the gas becomes ionized, and electrons and ions are accelerated by the electric field. When they hit other atoms, these are ionized as well, maintaining the ionization level. Atoms are excited by collisions, moving their electrons to higher energy levels. When an electron returns from a higher level to a ground state, a photon is emitted.

This process begins at the *starting voltage* (also known as the *striking voltage*, the *ignition voltage*, or the *breakdown voltage*) usually between 45V and 65V for standard types of bulb, or between 70V and 95V for high-brightness types.

When the bulb is operating, it emits a soft radiance known as a *glow discharge* with a wavelength ranging from 600 to 700 nanometers.

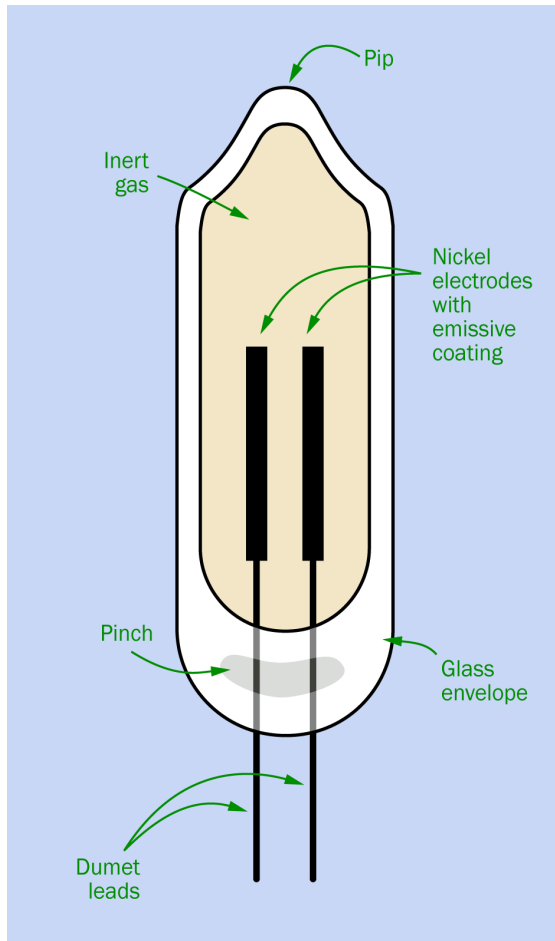


Figure 19-4. The parts of a neon bulb. See text for details.

The ionization of the gas allows current to flow through it. This will continue even if the power supply is reduced by 10 to 20 volts to a level known as the *maintaining voltage*.

Negative Resistance

When the glow discharge persists below the starting voltage, this is a form of *hysteresis*, meaning that the neon bulb tends to “stick” in its on state. It remains on while its power supply decreases to the maintaining voltage, but once it

switches off, it will “stick” in its off state until the power supply increases again above the maintaining voltage to the starting voltage. The concept of hysteresis is discussed in the entry on **comparators**. See [Figure 6-2](#).

A neon bulb is said to have *negative resistance*. If the current is allowed to increase without restraint, the resistance eventually decreases while the current increases further. If this runaway behavior is not controlled, the bulb will destroy itself.

This behavior is characteristic of gas-discharge tubes generally. A graph showing this appears in [Figure 19-5](#). Note that both scales are logarithmic. Also note that the curve shows how current will be measured in response to voltage. If the voltage is reduced after it has increased, the transitional events shown by the graph will not recur in reverse order. This is especially true if arcing is allowed to begin, as it will almost certainly destroy the component.

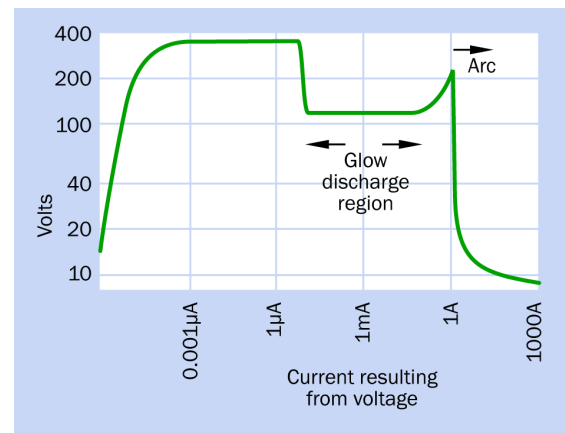


Figure 19-5. A gas discharge tube, such as a neon bulb, is said to have a negative resistance, as current passing through it tends to increase uncontrollably after the gas is ionized and becomes conductive. (Derived from measurements made by David Knight, on a web page named after his radio ham call sign, G3YNH.)

A neon bulb can be controlled very simply with a series resistor that maintains it in gas-discharge mode. To understand the operation of the resistor, consider the combination of the lamp and

the resistor as a voltage divider, as shown in [Figure 19-6](#). Before the lamp begins to pass current, it has an almost infinite resistance. Therefore, the voltage on both sides of the resistor will be approximately equal, the bulb passes almost no current, and it remains dark.

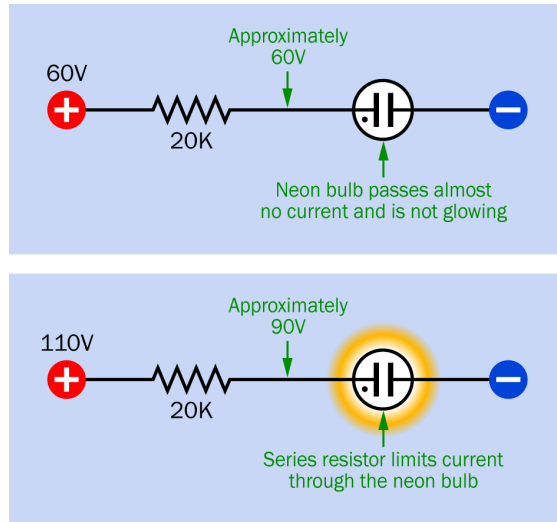


Figure 19-6. A series resistor is essential to limit the current through a neon bulb.

After the lamp begins to pass current, the requirement now is for the series resistor to reduce the voltage from the supply level (probably around 110V) to the maintaining level (probably around 90V). This means that the desired voltage drop is 20V, and if the manufacturer's specification tells us that the lamp should pass 1mA (i.e., 0.001 amps), R , the value of the series resistor, is given by Ohm's Law:

$$R = 20 / 0.001$$

Thus, the value for R is 20K. In fact, the value of a resistor supplied with a neon bulb may range from 10K to 220K, depending on the characteristics of the bulb and the supply voltage that will be used.

Now if the bulb's effective internal resistance falls radically, the resistor still limits the current. In a hypothetical worst-case scenario, if the bulb's resistance drops all the way to zero, the resistor

must now impose the full voltage drop of 110V, and the current, I , will be found by Ohm's law:

$$I = 110 / 20,000$$

That is, about 5mA, or 0.005A.

Neon tubes used in signage require a more sophisticated voltage control circuit which is not included in this encyclopedia.

How to Use It

The use of a neon bulb for an indicator lamp is primarily limited to situations where domestic supply voltage (115VAC or 220VAC) is readily available. "Power on" lights are the obvious application, especially as neon indicators can accept AC. The switch shown in [Figure 19-7](#) is illuminated by an internal neon bulb. The rectangular indicator in [Figure 19-8](#) is designed to run on domestic supply voltage, and its internal bulb and resistor can be clearly seen through the green plastic. The assembly in [Figure 19-9](#) is about 0.5" in diameter, which is the lower limit for neon indicators.



Figure 19-7. This power switch is illuminated by an internal neon bulb.

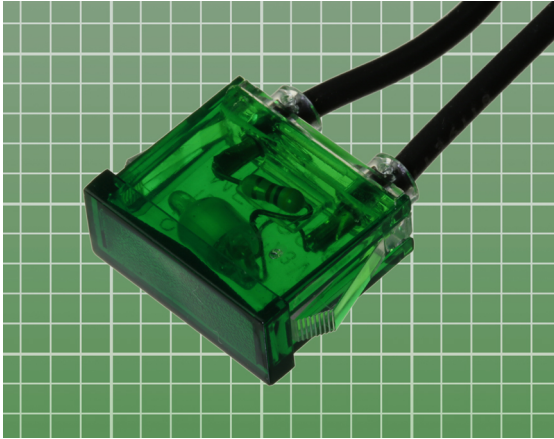


Figure 19-8. The neon bulb and its series resistor are visible inside this indicator.

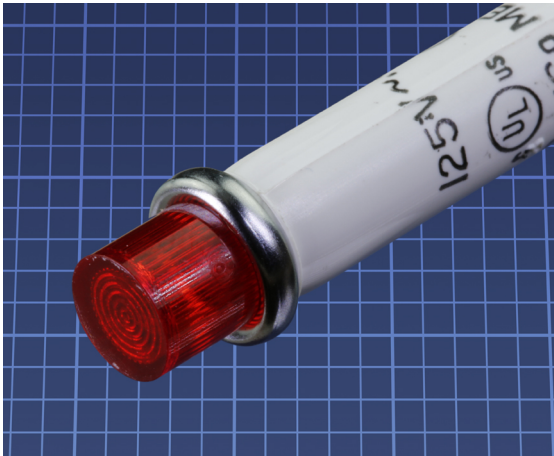


Figure 19-9. A relatively small neon indicator lamp, designed for insertion in a hole 0.5" diameter.

Limited Light Output

Neon bulbs have a light output of around 0.06 lumens per milliamp of consumed power (standard brightness type) or 0.15 lumens per milliamp of consumed power (high brightness type).

Comparing this value with the intensity of **LED** indicators is difficult. Their light output is customarily measured in *millicandelas* (mcd), because LED indicators almost always include a lens that focuses the light, and the candela is a measurement of luminous flux within an angle of dis-

persion. Moreover, because the intensity of neon indicators is not of great interest in most applications, datasheets usually do not supply an intensity value.

One way around the problem of comparisons is to use the standard of *radiant luminous efficacy* (LER), which is defined in the entry on incandescent lamps (see “Efficacy” on page 179). A standard-brightness neon bulb has an LER of about 50 lumens per emitted watt of luminous flux. A light-emitting diode may reach an LER of 100 lm/W. However, a neon bulb operates typically around 1mA while an LED indicator may use 20mA. Therefore, a typical LED indicator may appear to be 30 to 50 times brighter than a typical neon bulb.

Consequently, neon may be an inferior choice in a location where there is a high level of ambient light. Direct sunlight may render the glow of a neon indicator completely invisible.

Efficiency

Because a neon bulb does not use a lot of power and generates negligible heat, it is a good choice where current consumption is a consideration (for example, if an indicator is likely to be on for long periods). The durability and low wattage of neon bulbs, and their convenient compatibility with domestic power-supply voltage, made them a favorite for night-lights and novelty lamps in the past. [Figure 19-10](#) shows an antique bulb containing an ornamental electrode, while [Figure 19-11](#) is a piece of folk art, approximately 1" in diameter, mounted on a plug-in plastic capsule containing a neon bulb.

Ruggedness

Neon bulbs are a good choice in difficult environments, as they are not affected by vibration, sudden mechanical shock, voltage transients, or frequent power cycling. Their operating temperature range is typically from -40 to +150 degrees Celsius, although temperatures above 100 degrees will reduce the life of the lamp.



Figure 19-10. In bygone decades, ornamental neon bulbs with specially shaped electrodes were popular.



Figure 19-11. Neon folk art survives in this hand-painted night-light sold in a Florida tourist shop.

Power-Supply Testing

When driven by DC current, only the negative electrode (the *cathode*) of a neon bulb will glow.

When AC current passes through the bulb, both terminals will glow.

If a bulb (with series resistor) is placed between the “hot” side of a domestic AC power supply and ground, the bulb will glow. If it is placed between the neutral side of the supply and ground, it will not glow.

These features enable a neon bulb to be used for simple power-supply testing.

Life Expectancy

The metal of the electrodes gradually vaporizes during everyday use of a neon bulb. This is known as *sputtering* and can be observed as the glass capsule becomes darkened by deposition of vaporized metal. The electrodes will have a more limited life in a lamp used with DC voltage, where sputtering affects only the cathode. Using AC, the electrodes take turns functioning as the cathode, and vaporization is distributed between both of them.

Failure of a neon lamp can occur as sputtering erodes the electrodes to the point where the maintaining voltage will increase until it almost reaches the level of the power supply. At this point, the bulb will flicker erratically.

Failure can also be defined as a gradual reduction in brightness to 50% of rated light output, caused by accumulated deposition in the glass capsule. Because deposition occurs more heavily on the sides of the bulb, a longer apparent life is possible if the bulb is mounted so that it is viewed from the end.

Typically, neon bulbs are rated for 15,000 to 25,000 hours (two to three years of constant operation). However, the life can be greatly increased by a slight reduction in voltage, which may be achieved by substituting a series resistor with a slightly higher value.

The relationship between operating life and resistor value is shown below. If LA is the normal operating life, LB is the extended operating life,

RA is the normal resistor value, and RB is a higher resistor value:

$$LB = LA * (RB / RA)^{3.3}$$

For example, if a normal resistor value is 20K, and it is increased to 22K, the life of the lamp should increase by a factor of slightly more than 1.4.

Variants

A typical neon bulb terminates in leads, and a lamp assembly often has solder tabs, although it may have a base with a screw thread, flange, or bayonet pins for insertion into a compatible socket. A lamp assembly that does not use a base will either snap-fit into a hole of appropriate size and shape, or may be retained with a nut that engages with a plastic thread on the cylinder of the lamp.

Some neon bulbs or lamp assemblies terminate in pins for direct insertion into a printed circuit board.

Almost all neon bulbs operate either in the 100V to 120V range or in the 220V to 240V range.

Light intensity is expressed either as “standard” or “extra-bright,” although datasheets usually do not define those terms.

Nixie Tubes

Nixie tubes, first marketed in 1955, were used to display numerals from 0 through 9 in the days before LEDs took over this capability. They are no longer being manufactured.

Each numeral was physically formed from metal and functioned as an electrode inside a tube filled with a neon-based gas mixture. The typographical elegance of the digits and their aesthetically pleasing glow made Nixies enduringly popular. With a long lifespan, vintage tubes are still usable and can be purchased cheaply from sources such as eBay. Many originate in Russia, where Nixie-type displays were manufactured into the 1980s. The Russian tubes can be identi-

fied by their use of a numeral 5 that is a numeral 2 turned upside-down.

Nixie tubes typically require 170VDC. This creates a challenge for a power supply and switching, and can be a safety hazard.

Figure 19-12 shows six Nixie-type tubes repurposed for use as a 24-hour digital clock.

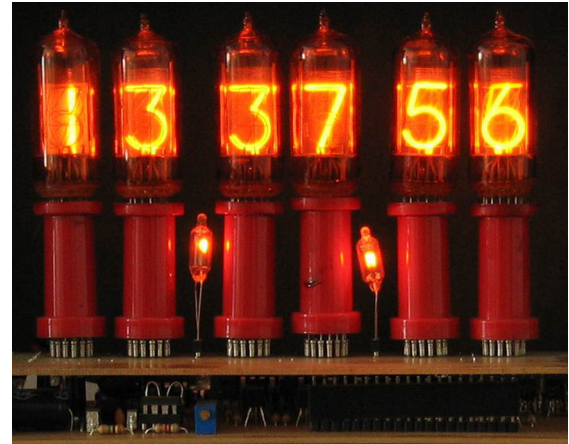


Figure 19-12. A 24-hour clock using Nixie-type tubes. Source: Wikipedia, public domain.

What Can Go Wrong

False Indication

Because a neon bulb requires so little power, it may be energized by induced voltages from elsewhere in a circuit, especially if inductive components such as transformers are used. To prevent this, a high-value resistor can be placed in parallel with the bulb, in addition to the series resistor that must always be used.

Failure in a Dark Environment

Because a neon bulb requires a minimal amount of light to initiate its own photon emissions, it may take time to start glowing in a very dim environment, and may not light at all in total darkness. A few bulbs include a small amount of radioactive material that enables them to self-start in complete absence of ambient light.

Premature Failure with DC

The life expectancy quoted in datasheets for neon bulbs usually assumes that they are powered by AC. Because DC results in faster vaporization of the electrodes, the expected lifetime should be reduced by 50% if DC power will be used.

Premature Failure through Voltage Fluctuations

Because the deterioration of a neon bulb accelerates rapidly with current, a sustained voltage

that passes slightly more current can radically reduce the expected lifespan.

Replacement

Replacement can be an issue with panel indicators, where disassembly of a device may be necessary to reach the bulb. Bear in mind, however, that an easily removable bulb becomes vulnerable to tampering.

fluorescent light

This entry deals primarily with *fluorescent tubes* (infrequently but sometimes described as *fluorescent lamps*), and *compact fluorescent lamps* (CFLs) that are marketed as a substitute for **incandescent lamps**. *Cold-cathode fluorescent lamps* (CCFLs) are also mentioned.

Vacuum fluorescent devices have a separate entry in this encyclopedia. A fluorescent tube or CFL does not contain a vacuum.

Although the diode(s) in a white **LED area lighting** unit are coated with a layer of fluorescent phosphors, they are not categorized here as fluorescent lights, and have their own entry.

A **neon bulb** resembles a fluorescent light in that it is a *gas-discharge* device, but the interior of its glass envelope is usually not coated with fluorescent phosphors, and therefore it has its own entry.

OTHER RELATED COMPONENTS

- **incandescent lamp** (see [Chapter 18](#))
- **LED area lighting** (see [Chapter 23](#))
- **vacuum fluorescent** devices (see [Chapter 25](#))
- **neon bulb** (see [Chapter 19](#))

What It Does

Fluorescent tubes or *compact fluorescent lamps* (CFLs) are primarily used for area lighting. A partially disassembled CFL appears in [Figure 20-1](#), showing the control electronics that are normally hidden inside the base.

There is no standardized schematic symbol to represent a fluorescent light. [Figure 20-2](#) shows three commonly used symbols for a fluorescent tube on the left, and three symbols for a CFL on the right. Note that two of the symbols for a CFL are the same as those for an **incandescent lamp**, shown in [Figure 18-1](#).

How It Works

Luminescence is the emission of light as a result of a process that does not require heat. (The opposite phenomenon is *incandescence*, in which heating causes an object to emit light; see [Chapter 18](#) for a description of **incandescent lamps**.)

Fluorescence is a form of luminescence. It occurs when electrons in a material are energized and then make a transition back to ground level, at which point they radiate their energy as visible light. The incoming energy can consist of other light at a higher frequency. Some creatures, including species of arachnids and fish, will fluoresce when they are lit with ultraviolet light.



Figure 20-1. A compact fluorescent lamp with its base cut away to reveal the control electronics.

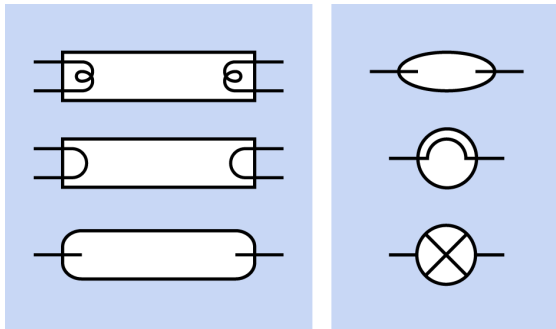


Figure 20-2. Schematic symbols to represent fluorescent tubes and bulbs are not standardized. See text for details.

A **fluorescent tube or lamp** contains a very small amount of mercury vapor that can be stimulated to emit ultraviolet light. This encounters a thin layer of **phosphors** coating the inner surface of the glass enclosure. The light causes the phosphors to **fluoresce**, emitting a diffuse radiance in the visible spectrum.

The tube or lamp also contains one or more inert gases such as argon, xenon, neon, or krypton at about 0.3% of normal atmospheric pressure. Two electrodes inside the glass enclosure are made primarily from tungsten, which can be preheated

to initiate ionization of the gas. Confusingly, both electrodes are often referred to as **cathodes**.

The function of the gas is not to emit light, but to conduct electric current, so that free electrons may encounter mercury atoms, raising their electrons briefly to a higher energy level. When one of these electrons reverts from its unstable energized state to its previous energy level, it emits a photon at an ultraviolet wavelength.

Figure 20-3 provides a diagram showing the interior of a fluorescent tube.

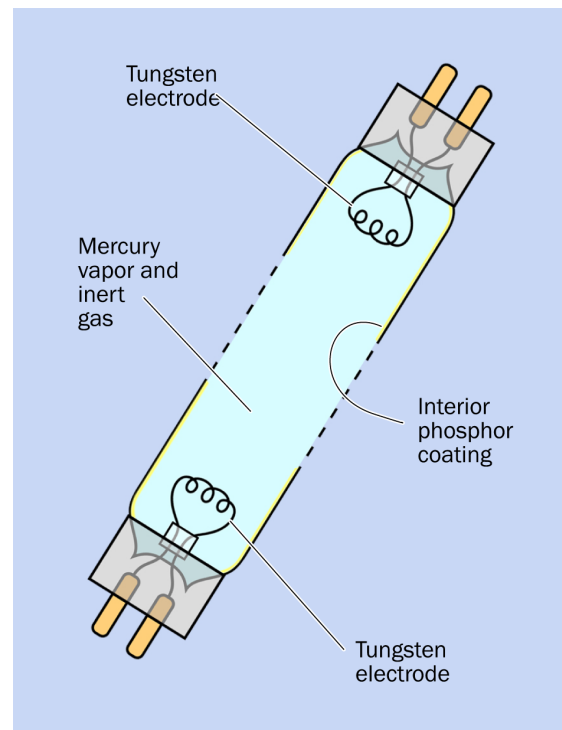


Figure 20-3. The basic parts of a fluorescent tube.

Ballast and Starter

Heating the tungsten electrodes is necessary but not sufficient to trigger ionization. A high-voltage pulse is also needed when the light is switched on. In a typical 48" tube, the pulse may range from 200V to 300V.

After current flow has been established, the gas, which is now a plasma, enters a phase of negative

resistance. Current passing through it will tend to increase even if the voltage decreases. This process must be controlled to prevent the formation of an *arc*, which will destroy the electrodes. (A similar process occurs in any gas discharge tube, such as a **neon bulb**, and is described in a graph in [Figure 19-5](#).)

To heat the electrodes, ionize the gas, and then control the current, the fixture for a fluorescent tube contains components that are separate from the tube. In their simplest, traditional form, these components consist of a *starter* and a *ballast*. The starter is a neon bulb that contains a bimetallic strip serving as a normally closed switch. It allows current to flow through the electrodes in series, to heat them. The basic circuit is shown in [Figure 20-4](#).

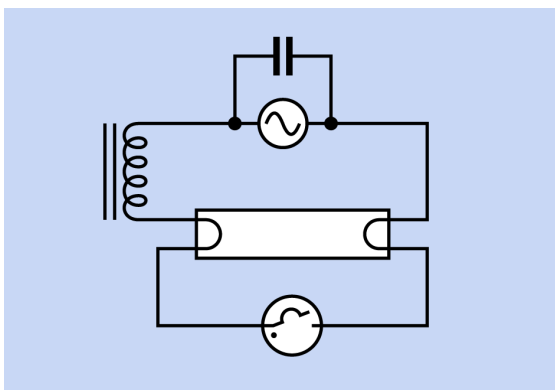


Figure 20-4. The traditional circuit to trigger ionization of the gas in a fluorescent tube uses a starter (shown at the bottom as a neon tube containing a bimetallic strip, which serves as a switch) and ballast (an inductive load, shown at left).

The starting process may not be immediately successful, in which case the starter may repeat several times in succession, causing the tube to flicker before its discharge becomes stable. In a cold environment, the tube will have more difficulty starting.

After the tube becomes conductive, current between the electrodes bypasses the starter. At this point, the *ballast* limits the current to prevent an

arc from forming. The simplest form of ballast is a coil that functions as an **inductor**.

In a more modern system, an *electronic ballast* replaces the starter-ballast combination. It not only applies the initial surge of high current but also raises the 50Hz or 60Hz frequency of the power supply to 10KHz or more. This increases the efficiency of the tube and eliminates any visible flickering of the light.

All compact fluorescent bulbs (CFLs) contain electronic ballasts. The small components visible in [Figure 20-1](#) are the ballast.

Flicker

When a fluorescent tube uses a conventional ballast and is illuminated with 50Hz or 60Hz AC, the glow discharge stops each time the current flow passes through the zero point in its cycle. In fact, the ionized gas in the tube cannot conduct until it is close to the maximum voltage, and stops conducting when the voltage rolls off. Consequently, the voltage across the tube fluctuates in an approximate square wave, and the light output begins and ends very abruptly. Although this occurs 100 times per second on a 50Hz supply and 120 times per second on a 60Hz supply, some people complain that the flicker is noticeable and can induce headaches.

The rapid on-off discharge is hazardous when it illuminates rotating parts in machinery, as a stroboscopic effect can make the parts seem to be stationary. To mitigate this effect, adjacent tubes in a fixture are powered by separate supplies that are out of phase. This is done either by using a three-phase power supply or by adding an LC circuit to the supply for one of the tubes.

Variants

The traditional type of ballast is also known as a *rapid-start ballast*. By preheating the electrodes, it reduces damage to them that otherwise tends to occur during the starting process. A tube designed for use with a rapid-start ballast has two

contacts at each end, and is referred to as a *bi-pin* tube.

An electronic ballast is also known as an *instant-start ballast*. It does not preheat the electrodes, and a tube designed to work with it has only one pin at each end.

CCFLs

A *cold cathode fluorescent lamp* (CCFL) may resemble a miniature fluorescent tube, typically measuring 2mm to 5mm in diameter. The tube may be straight or bent into a variety of shapes. It works on the same principle as a full-size fluorescent tube, containing mercury vapor and one or more inert gases, with an interior coating of phosphors to enable fluorescence. CCFLs are available in many colors and many shades of white.

As its name implies, the electrodes in a CCFL are not heated to establish ionization. Instead, a very high voltage (1,000VAC or more) is applied, dropping to 500VAC to 600VAC after the flow of current has been established. Because CCFLs have been often used to backlight laptop computer screens, inverter circuits are commonly available that create a high-frequency output at a high voltage from an input that can range from 3VDC to 20VDC. The inverter also includes provision to dim the CCFL by using *pulse-width modulation*.

Some CCFLs are designed for illumination of small spaces—for example, the interior of a display case. A few CCFLs look exactly like CFLs and can be used in light fixtures. Some may be compatible with the type of dimmer designed for incandescent lamps.

A CCFL usually has a limited light output compared with that of a conventional fluorescent tube, but has the advantage of working better at low temperatures. Some are designed for signage and exterior lighting in cold-weather locations.

They have a relatively long lifetime of up to 60,000 hours. A hot-cathode fluorescent lamp may fail between 3,000 and 15,000 hours.

Any tube or bulb that uses unheated electrodes to ionize a gas is technically a cold-cathode device, but will not be identified as a CCFL unless it also has an inner layer of phosphors to achieve fluorescence.

It is important to match a tube with the type of ballast installed in a fixture. This is not an issue with CFLs, as they have the appropriate ballast built in.

Sizes

Straight bi-pin tubes are sold in the United States in the following standard sizes:

- T5: 5/8" diameter. A more modern tube, but still with tungsten electrodes that serve to heat it.
- T8: 1" diameter. Very often 24" or 48" in length, consuming 18W or 36W respectively.
- T12: 1-1/2" diameter.
- T17: 2-1/8" diameter.

CFLs are sold in a very wide variety of configurations.

Comparisons

Fluorescent lights have significant advantages and disadvantages. On the plus side:

- After the fixture containing the ballast has been paid for, a tube is relatively cheap. A CFL or an LED light does not have this advantage, as the electronics are built in and will be discarded when the light fails.
- Fluorescent lights have a longer life than incandescent bulbs.
- Fluorescent lights are available in a wide range of shades of white.
- Fluorescent tubes create a diffuse radiance that is ideal for general lighting using ceiling-mounted fixtures. They do not cast harsh shadows.

On the minus side:

- Fluorescents were traditionally more energy-efficient than any other light source, but **LED area lighting** is now more efficient in some designs. LEDs are expected to become more efficient in the future.
- A fluorescent tube with a traditional type of ballast may cause complaints of flickering. By comparison, an LED light uses DC, and an incandescent bulb retains sufficient heat between power cycles so that it does not appear to flicker.
- Fluorescent flicker creates problems when shooting video.
- The fluorescent emission spectrum has sharp peaks that give the lighting an unnatural look.
- In applications that require a defined beam of light, a fluorescent source cannot be used.
- Conventional ballasts can create radio interference, especially in the AM band.
- Because fluorescent tubes and bulbs contain mercury, they require proper disposal, which can incur fees.
- Even an instant-on fluorescent light tends to hesitate briefly when it is switched on.
- The lifespan of a fluorescent light is greatly reduced if it is cycled on and off frequently. An incandescent bulb is less severely affected by cycling, and an LED light is not affected at all.
- Fluorescent lights have difficulty starting at low temperatures.

Values

Brightness

The intensity of a fluorescent light is measured in *lumens per watt*. Because invisible wavelengths are of little interest when assessing brightness, *luminous flux* is used to describe apparent bright-

ness in the visible spectrum. The unit for luminous flux is the *lumen*. Additional information about light measurement is included in the entry describing incandescent lamps (see “Power” on page 177).

Spectrum

The spectrum of photons emitted from mercury vapor in a fluorescent light has wavelengths that peak at 253.7 nanometers and 185 nanometers. (A *nanometer*, customarily abbreviated as nm, is one-billionth of a meter.) These wavelengths are invisible, being in the ultraviolet range, but when the light is transposed into the visible spectrum by the layer of phosphors, “spikes” in the range of wavelengths are still present. For a comparison of output curves for incandescent, fluorescent, and LED lights, see the graph in Figure 18-4.

Various formulations for the phosphors in a tube or CFL attempt to modify the character of the light to suit the human eye, but none of them looks as “natural” as the radiance from an incandescent bulb, probably because the characteristics of incandescent light are very similar to those of sunlight.

What Can Go Wrong

Unreliable Starting

At a low temperature, the mercury inside a fluorescent tube may be slow to vaporize. At very low temperatures, vaporization may not be possible at all. Until the mercury vaporizes, fluorescence will not occur.

Terminal Flicker

As a tube ages, it may start to conduct current only in one direction, causing it to flicker visibly. As it ages more, the gas discharge becomes even less reliable, and the flicker becomes erratic. Eventually, the gas discharge fails completely. In this state, a tube may show only a dim light at each end, in proximity to the tungsten electrodes.

Cannot Dim

Neither the older style of “conventional” ballast nor a modern electronic ballast will respond appropriately to a dimmer of the type designed for incandescent bulbs. This may be an important factor when an incandescent bulb is swapped out for a CFL.

Burned Out Electrodes

Like the tungsten filament in an **incandescent lamp**, the tungsten electrodes in a fluorescent tube suffer progressive erosion. This is evident

when a black tungsten deposit forms on the inside of the tube at one or both ends.

Ultraviolet Hazard

Some critics of CFLs maintain that the complex shape of a coiled or zig-zag tube tends to permit small imperfections in the internal phosphor coating, potentially allowing ultraviolet light to escape. If this occurs, and if a CFL is used in a desk fixture in close proximity to the user, ultraviolet light could elevate the risk of skin cancer.

laser

21

The term *maser* was coined in the 1950s to describe a device that used stimulated emission to amplify microwaves. When a device using similar principles amplified visible light in 1960, it was termed an *optical maser*. However, that term is now obsolete, having been replaced with **laser**. This term is always printed in lowercase letters, even though it is an acronym for Light Amplification by Stimulated Emission of Radiation.

The invented verb *to lase* is derived from *laser* and is used to describe the process of generating laser light, with the past participle *lased* and present participle *lasing* sometimes being used.

Thousands of laser variants exist. Because of space limitations, this entry will concentrate primarily on *laser diodes*, which are the smallest, most common, and most affordable type.

OTHER RELATED COMPONENTS

- **LED indicator** (see [Chapter 22](#))

What It Does

A **laser** generally emits a thin beam of intense light, often in the visible spectrum, and usually in such a narrow range of wavelengths, it can be considered *monochromatic*. The light is also *coherent*, as explained below.

Light output from a laser has three important attributes:

- **Intensity.** A high-powered laser can deliver energy to a very small, well-defined area, where it may be capable of burning, cutting, welding, or drilling. Large lasers may also be used as weapons, or for power transmission.
- **Collimation.** This term describes a beam of light that has parallel boundaries, and therefore does not disperse significantly when passing through a transparent medium such as air, glass, or a vacuum. A laser beam can have such excellent collimation, it can be

used in precision measuring devices, and has been transmitted over very long distances, even from the Earth to the Moon, where astronauts placed reflectors during the Apollo missions.

- **Controllability.** Because the beam can be generated with electrical power, its intensity can be modulated rapidly with relatively simple electronic circuits, enabling applications such as burning microscopic pits in the plastic of a CD-ROM or DVD.

Laser diodes are now more common than all other forms of lasers. They are found in pointers, printers, barcode readers, scanners, computer mice, fiber-optic communications, surveying tools, weapon sights, and directional lighting sources. They are also used as a light source to trigger more powerful lasers.

No generic symbol is used for a laser, but a laser diode is often represented with the same symbol

that is used for a light-emitting diode. See [Figure 22-2](#) in the entry for **LED indicators**.

How It Works

A laser is built around a *gain medium*, which is a material that can amplify light. The medium can be a solid, liquid, gas, or plasma, depending on the type of laser.

Initially, an input of energy provides stimulation for some atoms in the gain medium. This is known as *pumping* the laser. The energy input can come from a powerful external light source, or from an electric current.

Stimulation of an atom raises the quantum energy level of an electron associated with the atom. When the electron collapses back to its former energy state, it releases a photon. This is known as *spontaneous emission*.

If one of the photons encounters an atom that has just been excited by the external energy source, the atom may release two photons. This is known as *stimulated emission*. Beyond a threshold level, the number of released photons can increase at an exponential rate.

If two parallel reflectors are mounted at opposite ends of the gain medium, they form a *resonant cavity*. Light bounces to and fro between the reflectors, while pumping and stimulated emission amplifies the light during each pass. If one of the mirrors is partially transparent, some of the light will escape through it in the form of a laser beam. The partially transparent mirror is known as the *output coupler*.

Laser Diode

A laser diode contains an **LED**. (See [“How It Works” on page 207](#) for a more detailed description of the function of an LED.) The p-n junction of the diode functions as the resonant cavity of the laser. Forward bias injects charges into the junction, causing spontaneous emission of photons. The photons, in turn, cause other electrons and electron-holes to combine, creating more pho-

tons in the process of stimulated emission. When this process crosses a threshold level, current passing through the diode causes it to *lase*.

The original patent for a laser diode was filed by Robert N. Hall of General Electric in 1962, and the diagram in [Figure 21-1](#) is derived from the drawing in that patent, with color added for clarity.

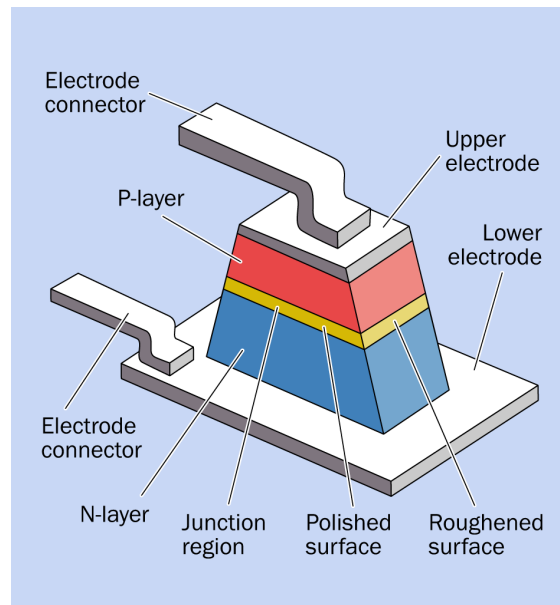


Figure 21-1. The original design for a laser diode, from the patent filed in 1962.

In the figure, the junction shown as a yellow layer forms the resonant cavity in which lasing occurs. It measures only 0.1 microns thick (the diagram is not drawn to scale). Its vertical front side is highly polished, and is parallel to the back side, which is also highly polished. Thus, photons reflect between these two vertical sides. The slanted face visible in the figure, and the other slanted face opposite it, are oriented and roughened to minimize internal reflection between them.

[Figure 21-2](#) shows a simplified cross-section of the laser diode.

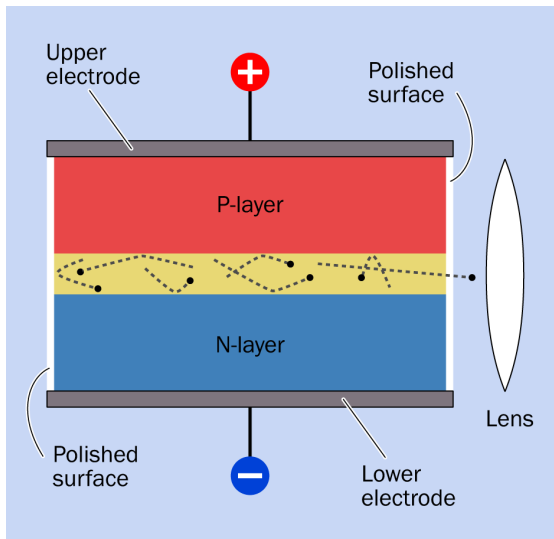


Figure 21-2. Simplified cross-section of a laser diode.

Figure 21-3 shows a cross-section of the diode installed in a component sold as a laser. It includes a photodiode to sense the intensity of light emerging through the polished rear end of the laser diode. External electronics are necessary to control the intensity of the laser, using feedback from the photodiode.

The component has three pins (shown pale yellow in the figure), one connecting to the photodiode, another connecting to the p-type layer of the laser diode, and the third being common to the n-type layer of the laser diode and the ground side of the photodiode.

A photograph of a laser diode is shown in Figure 21-4. Note the three pins, comparable to the pins shown in Figure 21-3, indicating that this component requires external control electronics.

In Figure 21-5, a laser is shown with a surface-mount chip adjacent to the solder pad connecting the blue wire. The presence of this chip, with only two wires, indicates that this component has its own control electronics and requires only a DC power supply.

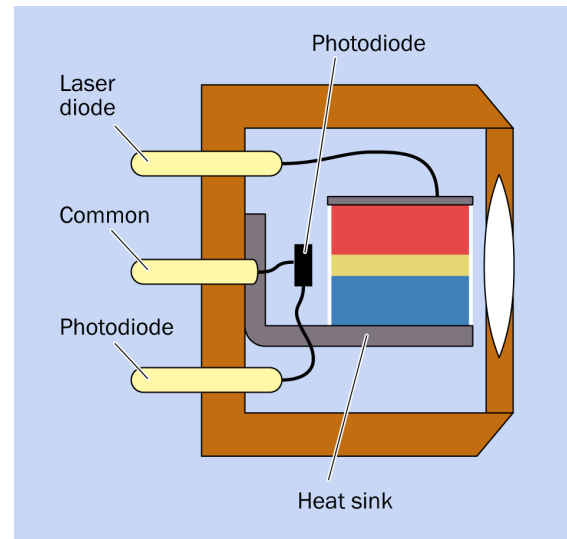


Figure 21-3. A laser diode is typically mounted with a photodiode to provide feedback for a driver circuit, to control the current consumed by the laser.



Figure 21-4. Lite-On 505T laser diode that emits light at 650nm. Power consumption 5mW at 2.6VDC. As indicated by the graph squares, this component is only about 0.2" in diameter.

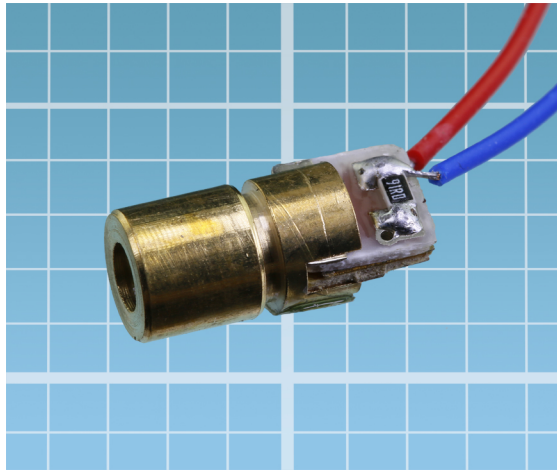


Figure 21-5. This laser incorporates its own control electronics and requires only a 5VDC power supply. It draws 30mA and generates an output up to 5mW.

Coherent Light

The emission of *coherent light* by a laser is often explained by suggesting that wavelengths are synchronized with each other. In fact, there are two forms of coherence that can be described approximately as *spatial coherence* and *wavelength coherence*.

If an observer looks up at a cloudy sky, the eye will perceive light radiating chaotically from many distances and directions. Thus, the light is not spatially coherent. The light also consists of many wavelengths, and thus it is not wavelength-coherent.

The filament of an incandescent lamp is a much smaller source of light, but still large enough to generate a profusion of light emissions that are spatially incoherent. The light also includes many different wavelengths.

Suppose a barrier containing a very tiny hole is placed in front of the **incandescent lamp**. If the aperture is very small, an observer on the far side will see the light as a point source. Consequently, the light that emerges from it is now spatially coherent, and will not have chaotically overlapping waves. If the light then passes through a filter, its wavelengths also will become coherent. This is

suggested in [Figure 21-6](#) where the light source is an incandescent lamp emitting a wide range of wavelengths.

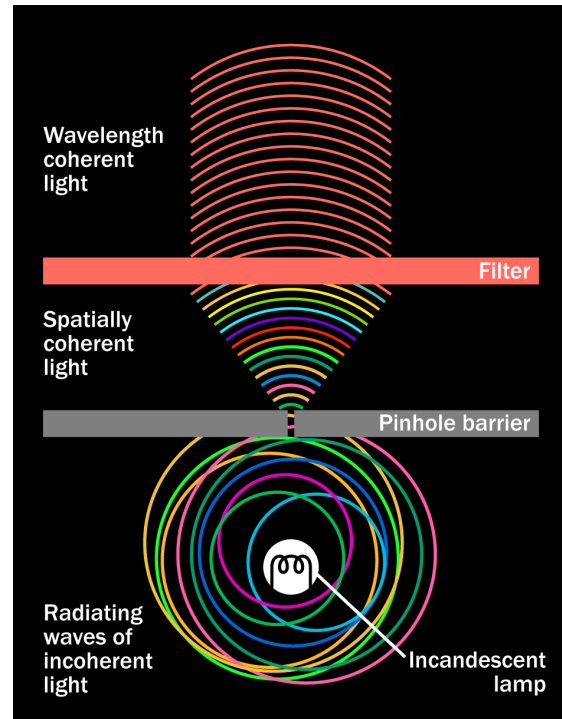


Figure 21-6. An incandescent lamp, at the bottom of the figure, emits incoherent light at many wavelengths (exaggerated here for clarity). When it passes through a pinhole, it becomes spatially coherent. When it then passes through a colored filter, it becomes wavelength coherent.

The small amount of light emerging through a pinhole is inevitably much dimmer than light from the original source. A laser, however, amplifies its light output, as well as tending to behave like a point source. The “hall of mirrors” effect of the parallel reflective surfaces in the resonant cavity causes much of the light to shuttle to and fro over a long distance before it emerges through the output coupler. Any light that deviates significantly from the axis of the laser will not escape at all, because the deviations will be cumulative with each reflection. Thus, the light from a laser appears to come from a point source at an almost infinite distance.

Because of the particular geometry of a light-emitting diode, the output from a laser diode is not naturally collimated, and tends to spread by an angle of around 20 degrees. A lens must be used to focus the beam.

Variants

Lasers are sold generally as fully assembled tools for a specific purpose. A very brief summary of CO₂, fiber, and crystal lasers is included here.

CO₂ Lasers

The gain medium is primarily carbon dioxide but also contains helium and nitrogen, with sometimes hydrogen, water vapor, and/or xenon. The laser is electrically pumped, causing a gas discharge. Nitrogen molecules are excited by the discharge and transfer their energy to the CO₂ molecules when colliding with them. Helium helps to return the nitrogen to base energy state and transfer heat from the gas mixture.

CO₂ lasers are infrared, and are commonly used in surgical procedures, including ophthalmology. Higher powered versions have industrial applications in cutting a very wide range of materials.

Fiber Lasers

Light is pumped via diodes and amplified in purpose-built glass fibers. The resulting beam has a very small diameter, providing a greater intensity than CO₂ lasers. It can be used for metal engraving and annealing, and also for working with plastics.

Crystal Lasers

Like fiber lasers, they are pumped by diodes. These compact lasers are available in a very wide variety of wavelengths, covering the whole visible spectrum, infrared, and ultraviolet. They find applications in holography, biomedicine, interferometry, semiconductor inspection, and material processing.

Values

The output power of a laser is measured in watts (or milliwatts). This should not be confused with the power consumed by the device.

In the United States, any device sold as a laser pointer is limited to a power output of 5mW. However, laser diodes packaged similarly to laser pointers can be mail-ordered with an output of 200mW or more. The legal status of these lasers may be affected by regulations that vary state by state.

In a CD-RW drive that is capable of burning a disc, the diode may have a power of around 30mW. A laser mounted in a CD-ROM assembly is shown in [Figure 21-7](#).

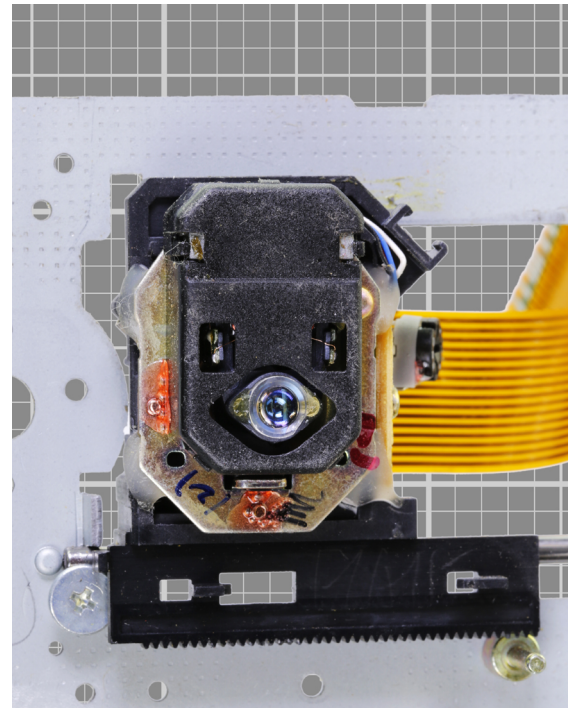


Figure 21-7. An assembly incorporating a laser for reading a CD-ROM.

Lasers have such a narrow range of wavelengths, they are given specific output values in nanometers. A laser in an optical mouse may have a wavelength of 848nm; in a CD drive, 785nm; in a

bar-code reader, 670nm; in a modern laser pointer, 640nm; in a Blu-ray disc player, 405nm.

How to Use It

While powerful lasers in a laboratory setting have exotic applications, a typical low-power laser diode has become so affordable (costing less than \$5 in some instances, at the time of writing) it can be considered merely as a useful source of a clearly defined light beam, ideal for detecting the position of a movable mechanical component or the presence of an intruder.

Generic light-emitting diodes are made with a view angle (i.e., a dispersion angle) as low as 3 degrees, but the beam is soft-edged compared with the precise boundary of a laser beam, and cannot be used reliably in conjunction with sensors at a distance of more than a few inches.

Laser diodes that are sold as components may or may not have current-limiting control electronics built in. Applying power to the laser diode directly will result in thermal runaway and rapid destruction of the component. Drivers for laser diodes are available separately as small, preassembled circuits on breakout boards.

For many applications, it may be simpler and cheaper to buy a laser diode as an off-the-shelf product. A laser pointer provides an easy way to get a source of laser light, and if it would normally be driven by two 1.5V batteries, it can be adapted to run off a 5V supply by using a 3.3V voltage regulator.

Common Applications

In addition to being used with PowerPoint presentations and in conjunction with position sensors, laser pointers have other applications:

- **Astronomy.** A high-powered laser beam is visible even in clear air as a result of interaction with air molecules. This is known as *Rayleigh scattering*. The phenomenon allows one person to point out a star (or planet) for another person. Because celestial objects are

so far away, parallax error is not detectable by two people viewing the beam while standing next to each other. A laser pointer may also be mounted on a telescope to assist in aiming the telescope at an object of interest. This is easier than searching for an object through an eyepiece.

- **Target acquisition.** Lasers are commonly used on firearms to assist in targeting, especially in low-light conditions. Infrared lasers can be used in conjunction with infrared viewing goggles.
- **Survival.** A small laser can be included in emergency supplies to signal search teams. A laser can also be used to repel predatory animals.

What Can Go Wrong

Risk of Injury

Lasers are potentially dangerous. Those that have an infrared or ultraviolet output are more dangerous than those with a visible beam, as there is no visual warning that the laser is active. A laser is capable of scarring the retina, although controversy exists regarding the power output that should be considered a high risk.

If a project incorporates a laser, it should be switched off while building or testing the device. It may be advisable to wear protective glasses that block laser light even when an experimenter feels confident that a laser is switched off.

Active lasers should never be pointed at people, vehicles, animals (other than dangerous animals), or oneself.

Inadequate Heat Sink

Lasers may be designed and rated for intermittent use. The burner assembly for a CD-ROM drive, for instance, will be rated for pulsed power, not continuous power. Read datasheets carefully, and provide an adequate heat sink.

Uncontrolled Power Supply

A diode laser that does not have a feedback system in place to control the flow of current can self-destruct.

aged by incorrect polarity of applied power. Pin functions should be checked carefully against datasheets.

Polarity

Both the light-emitting diode and the photodiode in a three-pin laser package can be dam-

LED indicator

In this encyclopedia, an **LED indicator** is defined as a component usually 10mm or smaller in diameter, made of transparent or translucent epoxy or silicone, most often containing one *light-emitting diode*. It is purposed as a status indicator in a device, rather than as a source of illumination, and is sometimes referred to as a *standard LED*.

LED indicators that emit infrared and ultraviolet light are included in this entry. LEDs that are designed to illuminate large interior or exterior areas are discussed in a separate entry as **LED area lighting**. They are sometimes described as *high-brightness LEDs* and almost always emit white light.

The term **light-emitting diode** is becoming less common, as the acronym **LED** has become ubiquitous. The acronym does not usually include periods between the letters.

The words “light emitting” are hyphenated here, as they form an adjectival phrase, but in everyday usage the hyphen is often omitted, and no definitive rule seems to exist.

Originally, a standard LED contained only one diode, but may now include multiple diodes, either to emit additional light or to provide a range of colors. In this encyclopedia, a single epoxy or silicone capsule is still considered to be an **LED indicator** regardless of how many diodes it contains. By contrast, any component consisting of multiple separately discernible light-emitting diodes, as in a seven-segment numeral, a 14- or 16-segment alphanumeric character, a dot-matrix character, or a display of multiple characters, is listed in a separate entry as an **LED display**.

OTHER RELATED COMPONENTS

- **LED area lighting** (see [Chapter 23](#))
- **LED display** (see [Chapter 24](#))
- **incandescent lamp** (see [Chapter 18](#))
- **neon bulb** (see [Chapter 19](#))
- **laser** (see [Chapter 21](#))

What It Does

An **LED indicator** emits light in response to a small current, typically around 20mA (but sometimes much less), at a voltage lower than 5VDC. It is usually molded from epoxy or silicone that may be colorless and transparent (often referred to as *water clear*), or colorless but translucent, or tinted and transparent, or tinted and translucent.

The color of the light is initially determined by the chemical compounds used internally, and by their dopants; therefore, a water-clear LED may emit colored light.

Ultraviolet LEDs are usually water-clear. Infrared LEDs often appear to be black, because they are opaque to the visible spectrum while being transparent to infrared.

When an LED indicator is described as being *through hole*, it has leads for insertion into holes in a circuit board. The term does not mean that the indicator itself is meant to be pushed through a hole in a panel, although this may also be done. The LED is cylindrical with a hemispherical top that acts as a lens. The leads are relatively thick, to conduct heat away from the component. A dimensioned diagram of a typical LED measuring 5mm in diameter is shown in Figure 22-1.

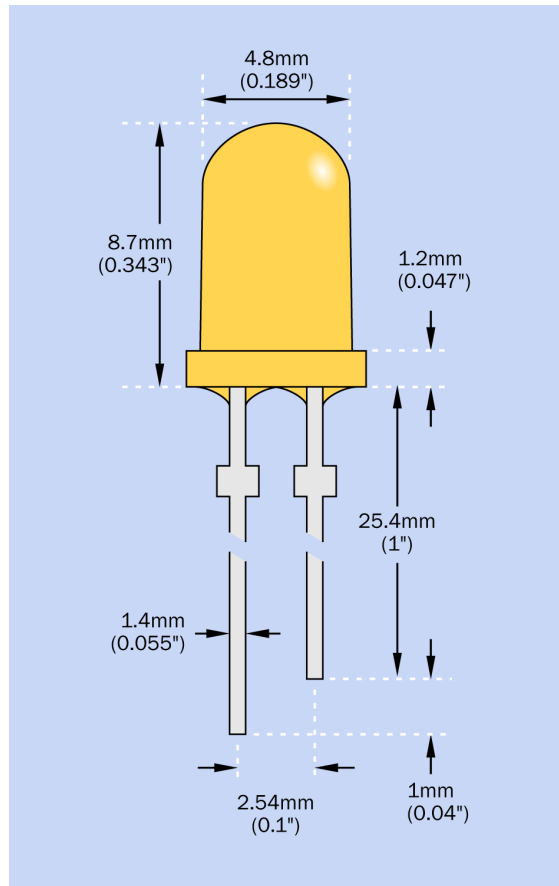


Figure 22-1. Dimensions of a typical 5mm LED. The longer wire connects with the anode, while the shorter wire connects with the cathode. Adapted from a data-sheet published by Lite-On Technology Corporation.

An LED indicator that is not the through-hole type is usually a surface-mount component. LEDs for surface mounting are mostly rectangu-

lar and can be as small as 1mm x 0.5mm. They may require a *heat sink*.

Schematic Symbols

Figure 22-2 shows a variety of symbols that are commonly used to represent an LED. The triangle at the center of each symbol points in the direction of conventional (positive-to-negative) current flow—from the anode to the cathode. Each pair of arrows radiating away from the diode indicates emitted light. Wavy arrows are sometimes used to represent infrared (thermal) radiation. Often, however, an infrared LED is represented in exactly the same style as an LED that emits visible light. With the exception of the wavy arrows, the various styles of schematic symbol are functionally identical and do not identify different attributes of the component such as size or color.

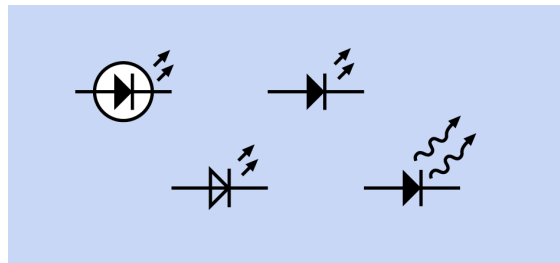


Figure 22-2. Various symbols may be used to represent an LED. See text for details.

Common Usage

LED indicators have mostly displaced **neon bulbs** and miniature **incandescent lamps** for the purpose of showing the status of a device. They are found in industrial control panels, home audio systems, battery chargers, washer/dryers, and many other consumer-electronics products. Higher output variants are used in flashlights, traffic signals, taillights on automobiles, and for illumination of subjects that are being photographed. LED indicators may be assembled in large numbers in attention-getting displays such as Christmas lights.

Red, orange, yellow, green, and blue are the basic standard colors. LEDs that appear to generate white light are common, but they do not emit an evenly weighted spectrum of wavelengths. See “Non-Incandescent Sources” on page 174 for a discussion of this topic.

How It Works

An LED, like any **diode**, contains a semiconductor *PN junction* that conducts current only in the forward direction (i.e., from the more-positive side of a power supply to the more-negative side). The diode becomes conductive above a *threshold voltage* sufficient to force electrons in the n-type region and holes in the p-type region to combine with each other. Each time this occurs, energy is released. The energy liberated by one electron-hole combination creates a *photon*, or one quantum of light.

The amount of energy released depends on the *band gap*, which is a property of the semiconductor material. The band gap is the smallest energy that can create an electron-hole pair. The energy determines the light’s wavelength, and thus the color.

The band gap also determines the threshold voltage of the LED. For this reason, LEDs of different colors have widely different threshold voltages.

Because an LED will often be used in devices where the DC power supply exceeds the maximum forward voltage, a *series resistor* is customarily used as a simple way to restrict current through the diode.

The light emitted by a colored LED indicator tends to include only a narrow range of wavelengths. However, the addition of a phosphor coating to the diode can broaden the output. This technique is used to make the light from a blue LED appear white, as shown in Figure 22-3. Most white LEDs are actually blue LEDs with a colored phosphor layer added. See the section

on **LED area lighting** in Chapter 23 for a more detailed discussion of this topic.

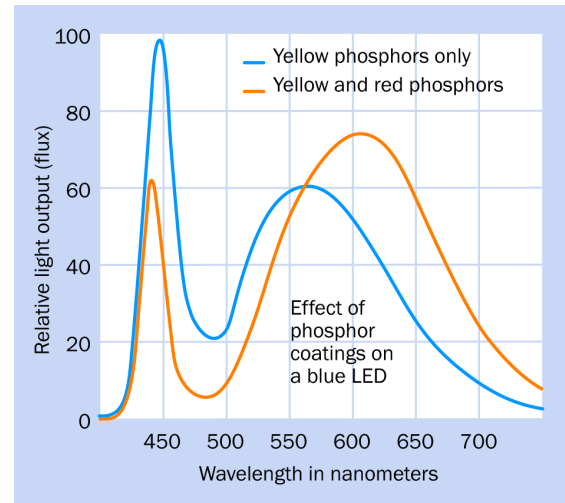


Figure 22-3. Increasing the range of emitted wavelengths by adding phosphors to a blue LED. Source: Philips Gardco Lighting.

Multicolor LEDs and Color Mixing

If red, green, and blue light sources are mounted extremely close together, the eye perceives them as a single source, of a color determined by their combined relative intensities. This system of additive color mixing is shown in Figure 17-14 in the entry dealing with LCDs. It is used in LED indicators that contain red, green, and blue light-emitting diodes in a single epoxy or silicone capsule.

While most video monitors use white LEDs or **fluorescent lights** to form a backlight for an **LCD** video screen, some high-end monitors use a matrix of very tiny red, green, and blue LEDs, because the combination of these separate colors generates a wider *gamut* of color wavelengths. The concept of gamut is discussed in “Color” in the **LCD** entry. The tiny LEDs in a backlight cannot be considered as indicators, but indicators are used for this purpose in billboard-sized video displays.

Variants

LED indicators vary widely in size, shape, intensity, view angle, diffusion of light, wavelength of light, minimum and maximum forward voltage, and minimum and maximum forward current.

Size and Shape

The original sizes for round LED indicators were 3mm, 5mm, or (more rarely) 10mm in diameter. Today, through-hole LEDs are sold in many intermediate sizes, although 3mm and 5mm are still most widely used.

The traditional round LED indicator is now augmented with square and rectangular shapes. In a parts catalog, a pair of dimensions such as 1mm × 5mm suggests that the LED is rectangular.

Intensity

The light intensity of an LED is usually expressed in *millicandelas*, abbreviated *mcd*. There are 1,000 mcd in a candela. For more information about units for measurement of light, see “Intensity” on page 178.

The candela measures the *luminous flux*, or visible radiant power, contained within a specified angle of dispersion, usually referred to as the *view angle*. This can be imagined as the rotated angle at the apex of a cone, where the cone defines the “spread” of the light, and the source is at the apex.

If a diode is emitting a fixed amount of luminous flux, the rating in mcd will increase with the inverse square of the view angle. This is because the light delivered to an area in front of the LED will become more intense as the angle becomes smaller. The use of mcd to rate the brightness of an LED can be misleading if it is not considered in comparison with the view angle.

For example, suppose an LED is rated at 1,000 mcd and has a view angle of 20 degrees. Now suppose the same diode is embedded in a different epoxy or silicone capsule with a lens that creates a view angle of only 10 degrees. The LED

will now be rated at 4,000 mcd, even though its total power output is unchanged.

- To compare the brightness of two LED indicators meaningfully, they should share the same view angle.

Four through-hole LED indicators with a wide range of specifications are shown in [Figure 22-4](#). From left to right: water-clear white generic, 10mm; Vishay TCR5800 5mm (emitting red, even though the capsule is water-clear), rated for 35,000mcd with 4 degrees view angle; Everlight HLMPK150 5mm red diffused, rated for 2mcd with 60 degrees view angle; and Chicago 4302F5-5V 3mm green, rated for 8mcd at 60 degrees view angle, containing its own series resistor to allow direct connection with a 5VDC power supply.

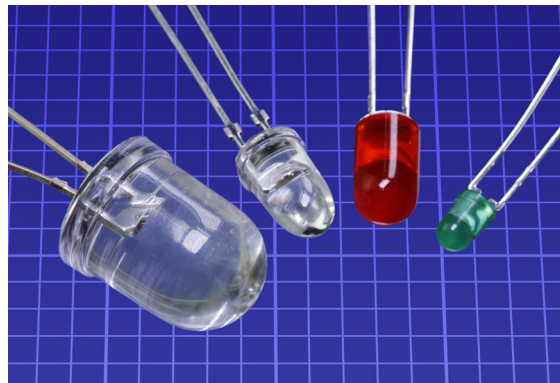


Figure 22-4. Four assorted LED indicators with very different specifications. See text for details.

Efficacy

The *radiant luminous efficacy* (LER) of an incandescent light source compares how effective it is at channeling its output within the visible spectrum, instead of wasting it in other wavelengths, especially infrared. Note that the word *efficacy* has a different meaning from *efficiency*. The LER acronym may help to avoid confusion.

LER is expressed in lumens per watt, and in an incandescent bulb it is calculated by dividing the power emitted in the visible spectrum (the *lumi-*

nous flux) by the power emitted over all wavelengths. This is described in detail on “Efficacy” on page 179 in the entry dealing with incandescent lamps.

In an LED indicator, almost all the radiation can be within the visible spectrum, but some power is wasted by generating heat internally. The efficacy varies depending on the type of LED; thus a red-orange indicator can have an efficacy of 98% while a blue LED will be probably below 40%.

Diffusion

Some LED indicators use epoxy or silicone that is formulated to be translucent or “cloudy” instead of transparent. They diffuse the light so that it is not projected in a defined beam, has a softer look, and has an approximately equal intensity when viewed from a wider range of angles.

“Clear” and “diffused” are options that must be taken into account when choosing LEDs from an online catalog, unless the user is willing to turn a clear LED into a diffuse LED by applying some sandpaper.

Wavelength and Color Temperature

The wavelength of light is measured in *nanometers* (abbreviated *nm*), a nanometer being 1 billionth of a meter. The visible spectrum extends from approximately 380nm to 740nm. Longer wavelengths are at the red end of the spectrum, while shorter wavelengths are at the blue end.

A typical LED emits a very narrow range of wavelengths. For example, Figure 22-5 shows the emission from a standard red LED indicator manufactured by Lite-On. Graphs of this type are typically included in manufacturers’ datasheets.

Because a red LED stimulates the cones in the eye that respond to red light, it “looks red” even though the color is not comparable with the natural red that is seen, for instance, in a sunset. That natural color actually contains an additional spread of wavelengths.

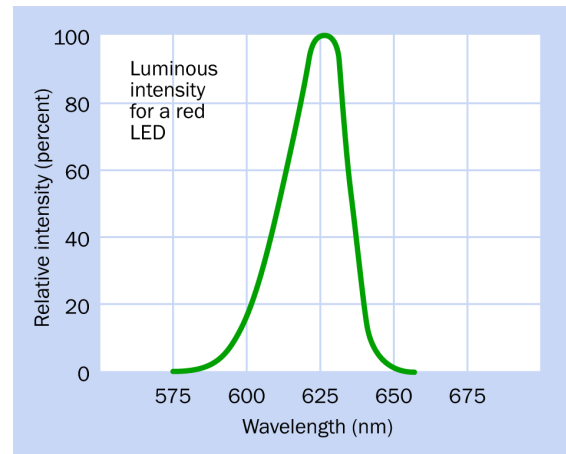


Figure 22-5. The narrow range of wavelengths emitted by a typical red LED indicator.

The following list shows the ranges of peak output values, in nanometers, for the most commonly available basic LED indicators (LEDs that emit other wavelengths are available, but they are less common):

- Infrared LED: 850 to 950
- Red LED: 621 to 700
- Orange LED: 605 to 620
- Amber LED: 590 to 591
- Yellow LED: 585 to 590
- Green LED: 527 to 570
- Blue LED: 470 to 475
- Ultraviolet LED: 385 to 405

Figure 22-6 shows this list graphically, omitting infrared and ultraviolet LEDs.

For almost 30 years, blue LEDs were a laboratory curiosity of little practical value, as efficiencies were stuck around 0.03%. An efficiency of more than 10% was finally achieved in 1995. Blue LEDs were marketed soon afterward.

However, when yellow phosphors are added to create the impression of white light by spreading the output over the whole visible spectrum, the wavelengths around 500nm are still not well represented, as suggested in Figure 22-3.

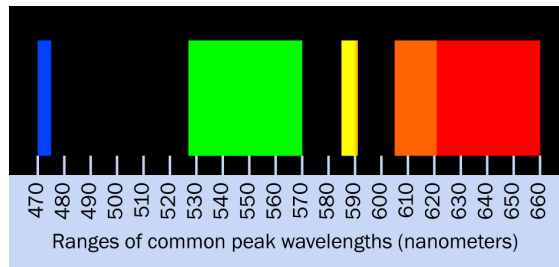


Figure 22-6. Ranges for peak wavelengths of the most commonly used LEDs. (Source: Survey of approximately 6,000 through-hole LEDs stocked at www.mouser.com.)

Fluorescent lights perform even more poorly than white LEDs, as can be seen in [Figure 18-4](#) in the entry describing **incandescent lamps**.

Because white LEDs do not emit a single peak of wavelengths, their color is expressed in color temperature rather than nanometers. The concept of color temperature is explained in “[Spectrum](#)” on page 173. White LEDs are available rated from 2,800 to 9,000 degrees Kelvin, and are discussed in more detail in the **LED area lighting** entry in this encyclopedia.

Internal Resistor

To eliminate the chore of adding a series resistor to limit current through an LED, some indicators are sold with a series resistor built in. They may be rated for use with 5VDC or 12VDC, but are externally indistinguishable from each other. They are also externally indistinguishable from LEDs that do not contain series resistors. [Figure 22-7](#) shows two 3mm LEDs, the one on the right containing its own series resistor, the one on the left being a generic LED without a series resistor.

Because of the nonlinear response of a diode, LEDs with or without internal resistors cannot be distinguished from each other reliably with a multimeter. If the meter is set to measure ohms, typically it will give an “out of range” error to all types of LED. If it is set to identify a diode, the reading will not tell you if the LED contains a resistor.

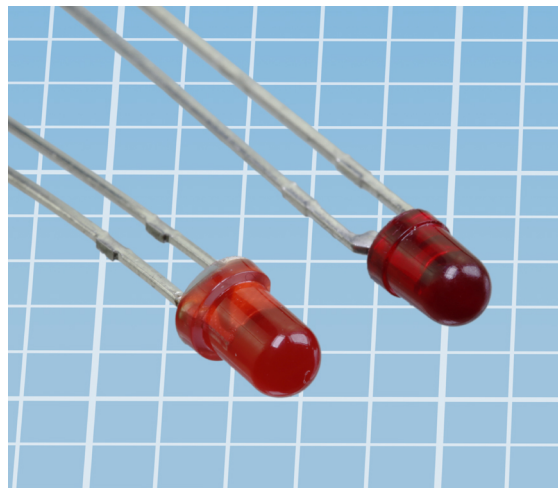


Figure 22-7. An LED (left) that does not contain its own series resistor is usually indistinguishable from one that does (right).

One way to determine whether an LED contains an internal series resistor would be to connect it with a variable power supply through a multimeter set to measure mA. Carefully increase the supply voltage from zero until the meter shows a current of 20mA. If the LED does not contain a series resistor, the supply voltage will be close to the recommended forward voltage for that type of LED (no lower than 1.6V for a red LED, and no higher than 3.6V for a white LED). If the LED does contain a series resistor, the supply voltage will be higher. This procedure is time consuming, but may be worthwhile to evaluate multiple LEDs that are known to be identical.

Multicolored

The leads for an LED indicator containing two or more diodes can be configured in several ways:

- Two leads, two colors. Two diodes are mounted internally in parallel, but with opposite polarity.
- Three leads, two colors. Two diodes share a common anode or common cathode.
- Four leads, three colors (RGB). Three diodes share a common anode or common cathode.

- Six leads, three colors. Three diodes, each with its own pair of leads, separate from the others.

Infrared

Most infrared emitters are LEDs that generate wavelengths longer than 800nm. They are found in handheld remotes to control consumer-electronics devices such as televisions and stereo systems, and are also used in some security systems, although *passive infrared motion detectors*, which assess infrared radiation from sources such as people or vehicles, are more commonly used for this purpose.

In conjunction with an infrared emitter, an infrared sensor is necessary, and must be sensitive to the same wavelength. To prevent false positives, the emitter modulates its output, typically with a carrier frequency between 10 and 100kHz. Remotes often use carrier frequencies of 30 to 56kHz. At the receiving end, the signal is processed with a band-pass filter matching the modulation frequency. Many different pulse-coding schemes are used, and no particular standard is dominant.

Ultraviolet

Because ultraviolet radiation can damage the eyes, LED indicators that emit ultraviolet light are potentially dangerous and should be used with caution. A yellow eyeshield can be worn to block the short wavelengths.

Ultraviolet light can be used to cure some adhesives and dental filling material. It can also kill bacteria, and can detect fluorescent print on bank notes, to check for counterfeiting. Ultraviolet flashlights are sold to detect some species of pests, such as scorpions, which fluoresce in response to ultraviolet light.

Values

The specification for an LED will include the wavelength of emitted light, luminous intensity, maximum forward voltage and current, maxi-

mum reverse voltage and current, and working values for voltage and current. All these values are important when choosing an indicator for a specific function.

White LEDs for room lighting or external use are calibrated differently. See the entry for **LED area lighting** in [Chapter 23](#).

Forward Current

About half of all the thousands of available types of LED indicators are rated for a typical forward current of 20mA to 25mA. Absolute maximum ratings may be twice as high, but should not generally be applied.

The light intensity of a typical 5mm red LED indicator is plotted against its forward current in [Figure 22-8](#). Note that current and light intensity have an approximately linear relationship up to the typical working current of 20mA. Even above this point, to the absolute maximum of 50mA, the light intensity rolls off only a very small amount.

Although an LED indicator can be dimmed by controlling the current passing through it, the current does not have a linear relationship with the applied voltage, and the indicator will stop functioning completely when the voltage drops below the threshold required by the diode. Consequently, LEDs are commonly dimmed by using *pulse-width modulation*.

Because of the nonlinear response of a diode, LEDs with or without internal resistors cannot be distinguished from each other reliably with a multimeter. If the meter is set to measure ohms, typically it will give an “out of range” error to all types of LED. If it is set to identify a diode, the reading will not tell you if the LED contains a resistor.

Low-Current LEDs

Indicators that require a very low forward current are convenient for direct connection to output pins of logic chips and other integrated circuits. Although a single output from an HC family chip

is capable of supplying 20mA without damaging the chip, the current will pull down the output voltage, so that it cannot be used reliably as an input to another chip while also lighting the LED.

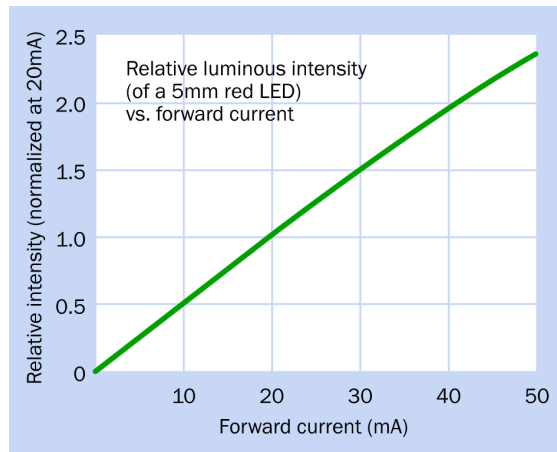


Figure 22-8. The relationship between forward current and light intensity of a typical 5mm LED indicator is approximately linear up to the 20mA operating current, and almost linear up to the absolute maximum of 50mA.

Various LED indicators drawing 2mA or 1mA are available, with intensities typically ranging between 1.5mcd and 2.5mcd. This low light output is still bright enough for viewing in a lab-bench environment. Low-current blue LEDs are not available. The only LEDs that draw as little as 1mA are red, as this is the most efficient type.

Using a higher value series resistor with a generic LED will of course reduce its current consumption, and some light will be visible so long as the forward voltage across the LED remains at its minimum level or above.

Forward Voltage

Red is the color that requires not only the least forward current, but the lowest forward voltage. In the range of 1.6VDC to 1.7VDC, all the LEDs are red. Typical forward voltages for various colors are shown here:

- Infrared LED: 1.6V to 2V
- Red LED: 1.6V to 2.1V

- Orange LED: 1.9V to 2.1V
- Amber LED: 2V to 2.1V
- Yellow LED: 2V to 2.4V
- Green LED: 2.4V to 3.4V
- Blue LED: 3.2V to 3.4V
- Ultraviolet LED: 3.3V to 3.7V
- White LED: 3.2V to 3.6V

Color Rendering Index

The *color rendering index* (CRI) evaluates the fidelity with which a light source is capable of displaying the full visible spectrum. It ranges from a perfect score of 100 down to 0 or even lower (sodium-vapor street lighting has a negative value). Computing the index requires standard reference color samples and has been criticized for generating scores that do not correlate well with subjective assessments.

Incandescent bulbs can have a CRI of 100, while an uncorrected white LED may score as low as 80.

Life Expectancy

Because the light output from an LED tends to decrease very gradually with time, the life expectancy is often defined as the number of hours required for the output to diminish to 70% of its output when new. Life expectancy is commonly stated on datasheets for high-brightness white LEDs, but is often omitted from datasheets for LED indicators.

Unlike incandescent lamps and fluorescent lights, LEDs do not have a shorter lifespan if they are frequently cycled on and off.

Light Output and Heat

The light intensity of an LED, measured in mcd, can vary from a few mcd to a maximum of 40,000mcd. Intensities above 30,000mcd generally are achieved by limiting the view angle to 15 degrees or less. Because the candela is weighted toward the central, green segment of the visible spectrum, green LEDs are likely to have a rela-

tively high mcd rating. LEDs rated between 20,000mcd and 30,000mcd, with a view angle of 30 degrees, are almost all green.

Datasheets may often include a *derating curve* showing the lower limit that should be placed on forward current through an LED indicator when its temperature increases. In Figure 22-9, the LED should be operated only within the boundary established by the green line.

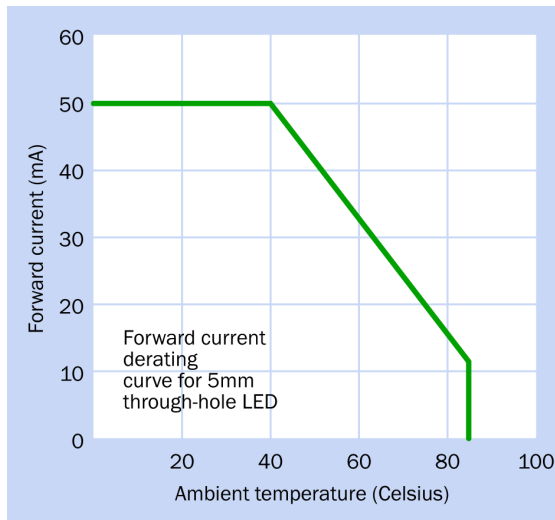


Figure 22-9. Safe operation of an LED entails limiting the forward current if the temperature increases. The green line shows the boundary for operation of this particular component.

View Angle

An LED formed from transparent epoxy or silicone (either water-clear or tinted) will create a well-defined beam with a view angle as narrow as 4 degrees or as wide as 160 degrees (in a few instances). The most common view angles for LED indicators are 30 degrees and 60 degrees.

Datasheets for LED indicators often include a *spatial distribution* graph showing the relative intensity of the light when viewed at various angles from the axis of the LED. The spatial distribution graph in Figure 22-10 is for an LED with a view angle defined as 40 degrees. This is the angle at

which the relative luminous intensity diminishes to 50%.

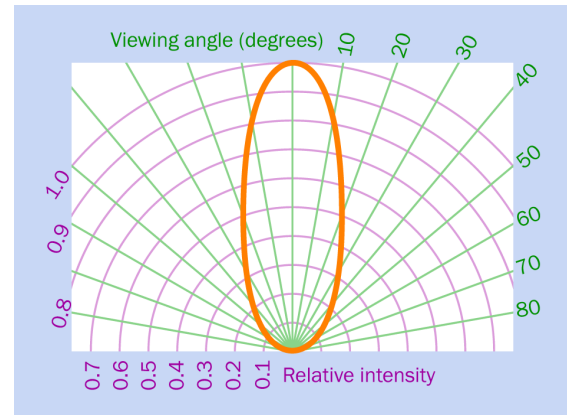


Figure 22-10. A spatial distribution graph shows the relative intensity of light from an LED at various view angles.

The view angle is of special concern in devices such as flashlights, where the spread of the beam affects the functionality.

How to Use It

Like all semiconductor devices, LEDs can be impaired by excess forward current and may break down irreversibly if subjected to excess reverse voltage. Their limits for reverse voltage are much lower than those of a rectifier diode. They are also vulnerable to heat, but are not particularly vulnerable to static electricity.

Polarity

A through-hole LED will have two leads of unequal length. The longer lead connects internally with the *anode* of the diode, and should be wired externally to the “more positive” side of a power source. The shorter lead connects internally with the *cathode* of the diode, and should be wired externally to the “more negative” side of a power source.

To remember the functions of the leads, consider that the plus sign would be twice as long as a minus sign if its horizontal and vertical stroke were disassembled and placed end to end.

If a round LED has a flange around its base, a flat spot in the flange will be closest to the cathode side of the component.

Series Resistor Value

Because the effective internal resistance of a diode is not a constant value at different voltages, a trial-and-error approach may be necessary to determine the ideal value for a series resistor with an LED indicator. For this purpose, a trimmer potentiometer can be used with a sample LED while measuring the current through it and the voltage drop across it. A fixed-value resistor can then be substituted. If the choice is between a resistor value that is a little too high and a value that is a little too low, the higher value resistor should be used.

An approximate value can be found using a very simple formula in which R is the resistor value, V_{CC} is the supply voltage, V_F is the forward voltage specified for the LED, and I is the desired current:

$$R = (V_{CC} - V_F) / I$$

Normally a series resistor rated at 1/4 watt will be acceptable, and 1/8 watt may be used in 5VDC circuits. However, care should be taken with a power supply of 9V or higher. Suppose an LED is rated for 1.8V forward voltage at 20mA. In a 5V circuit, the voltage drop across the series resistor will be:

$$V = 5 - 1.8 = 3.2$$

Therefore, the resistor must dissipate $3.2V \times 20mA = 64mW$. This is comfortably below the 125mW rating of a 1/8 watt resistor. However, with a 9V power supply, the voltage drop across the series resistor will be:

$$V = 9 - 1.8 = 7.2$$

Now the resistor must dissipate $7.2V \times 20mA = 144mW$. This exceeds the 125mW limit for a 1/8 watt resistor.

LEDs in Parallel

If multiple LEDs are to be driven in parallel, and none of them has to be switched individually, it is naturally tempting to save time by using a single series resistor for all of them. In these circumstances, assess the maximum current carefully and multiply by the voltage drop imposed by each of the LEDs, to determine the wattage of a series resistor.

Linking dissimilar LEDs in parallel is not recommended, because the threshold voltage decreases with increasing temperature. The hottest LED will therefore receive the largest current, and thus become even hotter. Thermal runaway can result.

LEDs containing their own series resistors can safely be wired in parallel.

Multiple Series LEDs

A series resistor wastes current by dissipating it as heat. In an application where two or more LED indicators will be illuminated simultaneously, the LEDs can be connected in series with a lower-value resistor, and three LEDs in series may eliminate the need for a resistor completely, depending on the voltage of the power supply. Here again a trimmer potentiometer should be used to determine an ideal value for any series resistor that may be necessary.

Comparisons with Other Light Emitters

Because LED indicators have largely replaced neon bulbs and miniature incandescent lamps, comparisons are of limited importance at this point. The situation regarding **LED area lighting** is different in that it is still competing actively with fluorescent lights and, in some instances, halogen. A list of advantages and disadvantages for high-intensity white LEDs is given in [“Comparisons” on page 223](#). The advantages of incandescent lamps are listed in [“Relative Advantages” on page 179](#).

Other Applications

LEDs are used in **optocouplers** and in **solid-state relays**. Usually an infrared LED is embedded inside a chip or a plastic module, and emits light through an interior channel to activate a phototransistor. This arrangement provides electrical isolation between the switching signal and the switched current.

Some sensors use an LED paired with a phototransistor at opposite sides of a U-shaped plastic mount. A sensor of this type can monitor industrial processes or may be found inside a photocopy machine, to detect the presence of a sheet of paper.

What Can Go Wrong

Excessive Forward Voltage

Like any diode, the LED has a *threshold voltage* in the forward direction. If this threshold is exceeded, the effective internal resistance of the LED falls very rapidly. Current rises equally rapidly, and quickly damages the component, unless it is protected by an appropriate series resistor.

Excessive Current and Heat

Exceeding the recommended value for forward current, or allowing an LED to overheat, will shorten its lifetime and cause a premature dimming of light output. LEDs generally require some current limiting or regulation (most commonly with a series resistor). They should not be connected directly to a voltage source such as a battery, even if the battery voltage matches the

voltage of the diode. The exception to this rule is if the internal resistance of the battery is high enough to limit the current, as in the case of button-cell batteries.

Storage Issues

LEDs of different types are often indistinguishable from each other. They can also be indistinguishable from photodiodes and phototransistors. Careful storage is mandatory, and reusing LEDs that have been breadboarded may cause future problems if they are wrongly identified.

Polarity

If the leads on an LED indicator are trimmed, and if the indicator lacks a flange in which a flat spot will identify the cathode, the component is easily misused with reversed polarity. If it is connected with a component that has limited current sourcing capability (for instance, the output pin of a digital chip), the LED will probably survive this treatment. However, maximum reverse voltage is often as low as 5VDC. To minimize the risk of errors, the anode lead can be left slightly longer than the cathode lead when they are trimmed for insertion in a breadboard or perforated board.

Internal Resistors

As previously noted, it is difficult to distinguish an LED that contains its own series resistor from another LED that does not. The two types should be stored separately, and should be reused circumspectly.

LED area lighting

The term **LED area lighting** is used in this encyclopedia to describe a white LED source that is bright enough to illuminate rooms, offices, or outdoor areas. It may also be used in desk lamps or table lamps as *task lighting*. LEDs for these purposes may be categorized as *high-brightness*, *high-power*, *high-output*, or *high-intensity*. A complete fixture containing at least one light source is properly known as a *luminaire*, although the term is not uniformly applied and is sometimes written incorrectly as a *luminary*.

The full term *light-emitting diode* is not normally applied to an LED used for area lighting. For this purpose, the LED acronym has become universal. Periods are not normally placed between the letters.

While an LED area-lighting package may contain more than one diode, it is still categorized here as a single source. By contrast, any component consisting of multiple separately discernible light-emitting diodes, as in a seven-segment numeral, a 14- and 16-segment alphanumeric character, a dot-matrix character, or a display of multiple characters, is listed in a separate entry as an **LED display**.

The term *OLED* is an acronym for *Organic Light-Emitting Diode*, a thin panel in which an organic compound is contained between two flat electrodes. Despite its functionality as a form of LED, its design is similar to that of thin-film electroluminescent light sources. Therefore it is discussed in the entry on **electroluminescence**.

OTHER RELATED COMPONENTS

- **LED indicator** (see [Chapter 22](#))
- **incandescent lamp** (see [Chapter 18](#))
- **fluorescent light** (see [Chapter 20](#))
- **neon bulb** (see [Chapter 19](#))
- **electroluminescence** (see [Chapter 26](#))

What It Does

High-brightness white LEDs provide a plug-compatible alternative to **incandescent lamps**, *halogen lighting*, and **fluorescent lights** for work spaces and the home.

At the time of writing, products are still evolving rapidly in the field of LED area lighting. A shared goal of manufacturers is to increase efficiency while reducing retail price to the point where

high-brightness LEDs will displace fluorescent tubes for most low-cost lighting applications.

A wall-mounted LED reflector-bulb that emulates a halogen fixture is shown in [Figure 23-1](#). A small LED floodlight for exterior use is shown in [Figure 23-2](#). An early attempt to package an LED area light in a traditional-style bulb is shown in [Figure 23-3](#). Within a decade, as LED area lighting continues to evolve, some of these examples

may look quaint. Configurations are evolving, with final results that remain to be seen.



Figure 23-1. A small LED reflector-light emulating a halogen fixture. Note the square of yellow phosphors mounted on the diode.

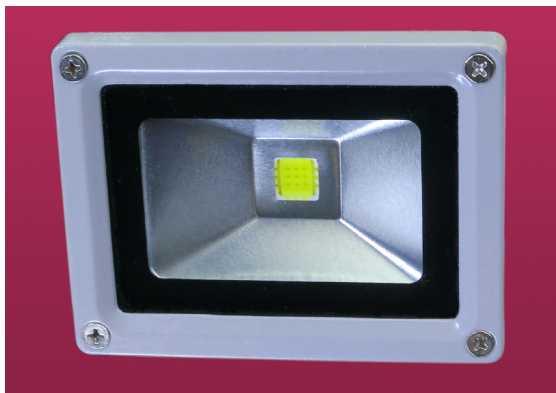


Figure 23-2. A floodlight for exterior use. Nine LEDs are assembled behind the yellow phosphors. The steel frame measures about 4" by 3".

Trends in Cost and Efficiency

The *luminous flux* of a source is the total power that it radiates in all directions, within the visible spectrum. The unit that measures luminous flux is the *lumen*. For a more detailed discussion of

this topic, see [“Power” on page 177](#). Since 1965, the cost per lumen of light from a given color of LED has decreased by about a factor of 10, while the maximum number of lumens emitted by an LED package has increased by a factor of 20, during each decade. This is known as *Haitz’s Law*, named after Dr. Roland Haitz of Agilent Technologies. [Figure 23-4](#) illustrates it graphically.



Figure 23-3. An LED light bulb. Unlike an incandescent bulb, it focuses the illumination in one direction, like a reflector-light. Consuming only 6W, it is claimed to be equivalent to a 40W incandescent bulb.

Schematic Symbol

Schematic symbols that are commonly used to represent an LED are shown in [Figure 23-5](#). The symbol remains the same regardless of the size or power of the component, but architectural plans may represent any type of light using the circle-and-X symbol at bottom right.

How It Works

A high-brightness LED functions on the same basis explained in the entry describing **LED indicators**. Photons are emitted when electrons are sufficiently energized to cross a PN junction and combine with electron-holes.

An LED that appears white, or off-white, actually emits blue light that is re-radiated over a wide range of wavelengths by adding a layer of yellow phosphors to the chip. A cutaway diagram of an

LED chip (properly known as a *die*), mounted under a silicone lens, is shown in [Figure 23-6](#).

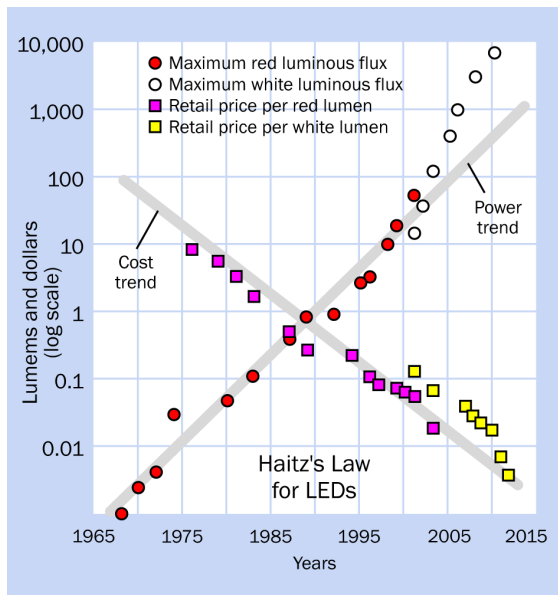


Figure 23-4. The increase in light output (luminous flux, in lumens) of a single LED compared with the decrease in cost-per-lumen during the years since 1965. The vertical logarithmic scale measures both dollars and lumens. Source: Philips Gardco site-lighting fact sheet with additional data from a “Strategies in Light Report” published by Semiconductor Equipment and Materials International in 2013.

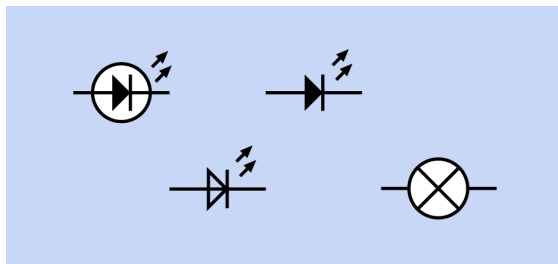


Figure 23-5. The symbol for an LED remains the same regardless of its size and power, but architectural plans may use the circle-and-X symbol at bottom right for any type of light fixture.

LEDs are mass-produced by etching them into crystals that are then cut into *wafers* before being subdivided into *dies*, like silicon chips. Most of the blue LEDs that form the basis of white lighting

use sapphire crystals as their substrate. The crystal may range in diameter from two inches to six inches. Large sapphire wafers are also finding potential applications in camera lens covers and scratch-resistant cover plates for cellular phones.

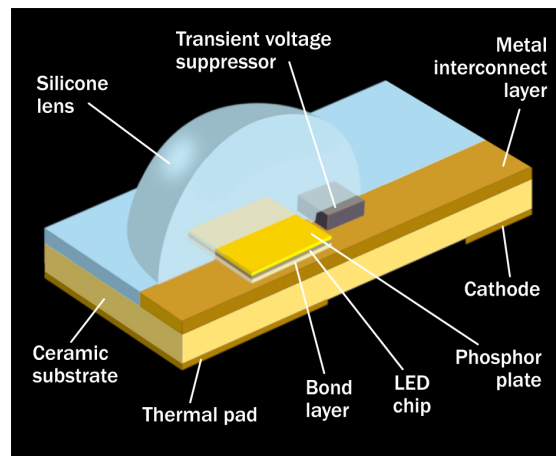


Figure 23-6. Cutaway diagram of a high-brightness white LED. Adapted from Philips Lumileds Technical Reference document.

While a die for an LED indicator may be 0.3mm x 0.3mm, a die in a high-brightness LED is often 1mm x 1mm. The size is limited by technical issues involving total internal reflection of the generated light.

The precise color of the light is adjusted by adding red phosphors to the yellow phosphors. This reduces the overall efficiency of the LED by around 10% but creates a “warmer” radiance. A graphical illustration of this principle is shown in [Figure 22-3](#), in the entry on LED indicators.

The *color temperature* of white or offwhite light is measured in degrees Kelvin, typically ranging from 2,500K to 6,500K, where a lower number represents a light with more red in it and a higher number represents a light with more blue in it. This system of measurement was originally used with **incandescent bulbs** to define the temperature of the filament, which determined its color. See “[Spectrum](#)” on page 173 for a detailed explanation.

Visible Differences

The effects of different types of illumination are compared in [Figure 23-7](#). To create this figure, first a color chart was prepared in Photoshop and printed on high-gloss photo-grade paper with a Canon Pro9000 Mark II inkjet printer, which has separate colors for red and green in addition to cyan, pale cyan, magenta, pale magenta, yellow, and black.

The color chart was then photographed twice with a Canon 5D Mark II, using a fixed white balance of 4000K. The first exposure was made with “daylight spectrum” LED lighting (claimed color temperature of 6500K) while the second was made with halogen lighting (claimed color temperature of 2900K). The photographs were adjusted in Photoshop for levels only, to fill the available range of 256 values. The two exposures show how the same chart would appear when viewed under the different lights, if the human eye did not adjust itself at all. Note the large area of the LED exposure which is rendered in shades of blue or purple. Also note the dullness of the reds. This confirms the everyday belief that “daylight spectrum” LEDs tend to have a cold, purplish cast while incandescents have a warmer, yellow look.

The same camera was then used to make two more exposures, this time with the white balance set to 6500 for LED lighting and 2900 for halogen lighting, which would be the recommended standard procedure, suggesting the kind of compensation that the human eye also tends to make for different ambient lighting. The result is shown at [Figure 23-8](#). The LED version has improved, but the reds and yellows are still muted. The halogen version also looks better than before, but the magenta end of the spectrum has too much yellow in it. These images show the limits of white-balance correction for indoor photography.

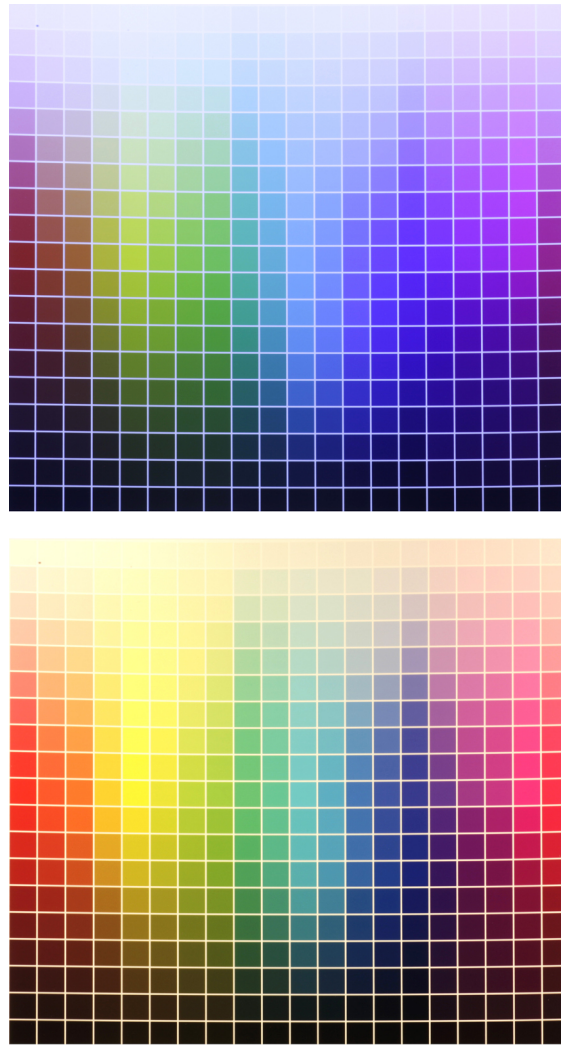


Figure 23-7. The same printed color chart viewed with “daylight white” LEDs (top) and halogen lighting (bottom), without any compensation to allow for the different light spectra. A fixed white balance of 4000K was used for both pictures.

Side-by-Side Comparison

Because the human eye is much better at comparing colors when they are adjacent to each other, another color chart was prepared using just six color bars of fully saturated red, yellow, green, cyan, blue, and magenta, with paler and darker versions above and below. The bars were separated with large white gaps. The chart was photographed first with the white balance set to

6500 under “daylight white” LED lighting and then again with the white balance at 2900 for halogen lighting. In Photoshop, the bars from the halogen version were copied and pasted beside the bars from the LED version to facilitate an A-B comparison. The result is shown in [Figure 23-9](#).

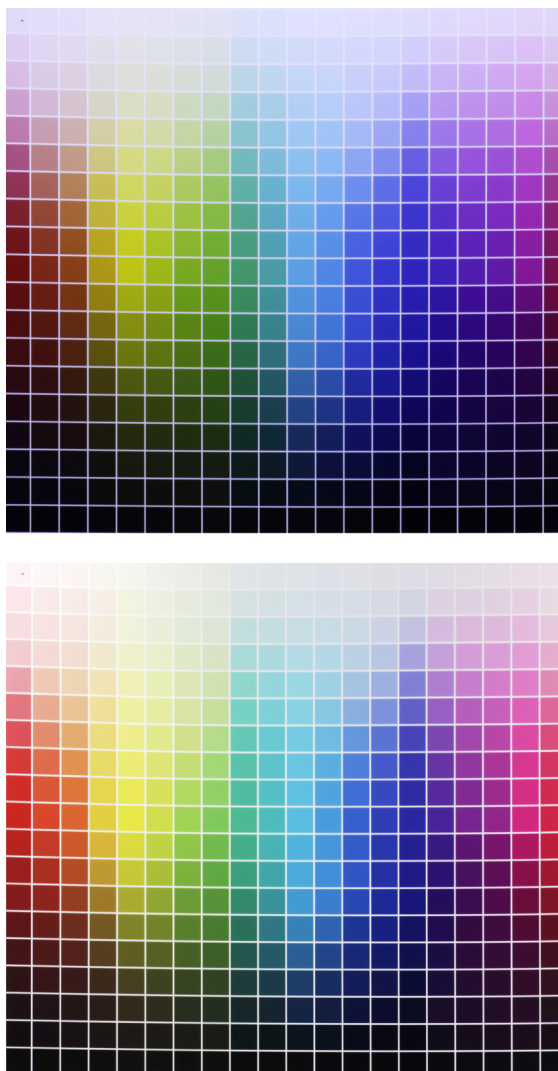


Figure 23-8. The same color chart as before, photographed with appropriate color temperature settings of 6500K (top, using “daylight-spectrum” LEDs) and 2900K (bottom, using halogen).

In each pair of colors, LEDs lit the one on the left, and halogen lights lit the one on the right. This

figure shows the dramatic difference at the red end of the spectrum, and the poor reproduction of yellows by LED lighting. However, the LED rendering of green is better, and likewise the rendering of magenta, except where its darker version is concerned. Among the pale versions of the colors, the LED lights produce much less density (i.e., they have a brighter look) in the blues, greens, and cyans. The low densities will show up as pale highlights in a photograph of an object, and the picture will tend to have excessive contrast. This will also contribute to the “harsh” look of “daylight white” LED lighting which may be perceived by the eye.



Figure 23-9. A range of six fully saturated colors, with lighter and darker shades added above and below, photographed first with “daylight spectrum” LEDs and then with halogen lights, after which the two sets of colors were paired for easy comparison. The LED version is on the left in each pair.

Halogen is deficient at the blue-violet end of the spectrum, even when the camera has an appropriate white-balance setting. Photographers can correct this using image-editing software. LED “daylight spectrum” lights are more difficult to correct. LEDs classified as “warm” should reproduce reds better, but may not do so well with blues.

Diffuse light from a uniformly cloudy sky may be the most ideal form of lighting for photographing objects, but this is of little help for people who work (or take photographs) under artificial lights.

LED lights that contain separate red, green, and blue emitters may perform better, but create a different problem, in that shadows will tend to have color fringes caused by the small offsets between the colored emitters.

Heat Dissipation

An LED is less than 100% efficient because not all electrons mate with electron holes. Some manage to bypass the semiconductor junction; some recombine without generating light; and some transfer their energy to other atoms. In each instance, waste heat is created. While the heat in an incandescent bulb is mostly dissipated by radiation, an LED must get rid of the heat almost entirely by conduction, typically through a *heat sink*. This complicates the design of a fixture, because the integrity of the pathway to dispose of the heat must be retained when the LED bulb or tube is replaced.

Efficacy

The *radiant luminous efficacy* (LER) of an incandescent light source measures how effective it is at channeling its output within the visible spectrum instead of wasting it in infrared radiation. LER is expressed in lumens per watt, and in an incandescent bulb, it is calculated by dividing the power emitted in the visible spectrum (the *luminous flux*) by the power emitted over all wavelengths. This is described in detail in “Efficacy” on page 179 in the entry dealing with incandescent lamps.

In an LED indicator, almost all the radiation can be contained within the visible spectrum, which suggests that its efficacy should be 100%. However, because some waste heat is still created internally, the efficacy is calculated by dividing the light output, in lumens, by the power input, in watts, at the voltage required by the LED. (Lumens can be converted directly to watts, and therefore this division makes a comparison between similar units).

In an LED lighting fixture that contains its own electronics to convert higher voltage AC to lower

voltage DC, the power consumption of the fixture is measured not at the diode, but at the input side of the electronics. Therefore, the inefficiency of the electronics reduces the efficacy value of the lighting unit.

Dimming

An incandescent bulb is very sensitive to reduction in power. It becomes radically inefficient, emitting perhaps 1% of its normal light output if the power is reduced to 40%.

LEDs have an almost linear response to the supplied power. Usually a triac-based dimmer will not work well with LED area lighting, and a dimmer designed for LEDs must be substituted, using pulse-width modulation.

Ultraviolet Output

The gas plasma in a fluorescent light generates ultraviolet wavelengths that are shifted to the visible spectrum by the phosphor coating inside the glass envelope. Imperfections in the phosphor coating can potentially allow leakage of ultraviolet light, causing some researchers to claim that the use of CFLs (compact fluorescent lighting) for close-up work with desk lamps can increase the risk of developing some forms of skin cancer. (This claim remains controversial.)

LED manufacturers are quick to point out that white LEDs do not emit any ultraviolet radiation. Figure 23-10 shows spectral power distribution curves derived from measurements of three high-brightness Color Kinetics LEDs manufactured by Philips. The manufacturer states categorically that “The LED-based color and white light products made by Color Kinetics do not emit outside the visible spectrum.” Infrared radiation is also negligible.

Color Variation

The *correlated color temperature* (CCT) is determined by finding the conventional incandescent color temperature which looks most similar to the light from a white LED. Unfortunately, because the CCT standard is insufficiently precise,

and because small manufacturing inaccuracies can occur, two LED sources with the same CCT number may still appear different when they are side by side. While the human eye adjusts itself to overall color temperature, it is sensitive to differences between adjacent sources. If two or more white LEDs in a lighting fixture do not have identical spectra, the difference will be noticeable.

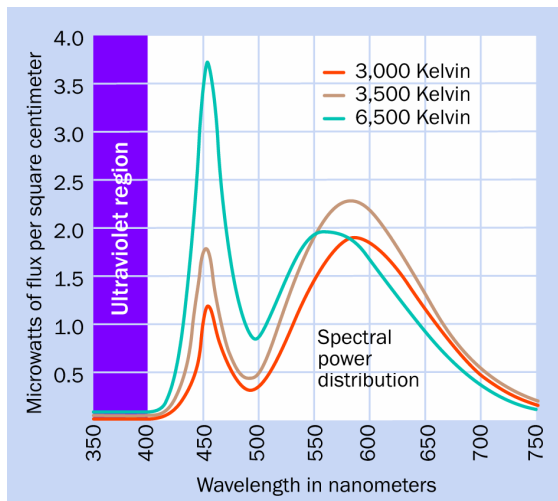


Figure 23-10. Spectral power distribution curves for three high-brightness white LED lamps showing no ultraviolet emissions. (Adapted from a graph in a white paper published by Color Kinetics Incorporated.)

To address the issue, manufacturers introduced the concept of “binning,” in which lights are subclassified to tighter specifications and are assigned bin numbers based on their measured characteristics. The Philips Optibin system, for instance, assesses the light from various angles, as well as perpendicularly to the source. This is especially important where a large area such as a building lobby is painted white and is lit by high-brightness LEDs that must appear uniform in color temperature.

Variants

LED area lighting products are often designed to emulate the form factors of incandescent bulbs,

halogen reflector bulbs, or fluorescent tubes. The standard screw-in base of an LED light bulb, the pin-base of a miniature 12V LED reflector bulb, and the pins on an LED tube enable easy migration to the newer technology.

Strip lights are unique to LED lighting systems. They are thick, flexible plastic ribbons in which are embedded a series of LEDs. For area lighting, the LEDs are white, and the strips can contain necessary control electronics for conversion of AC power. The strips can be placed behind ledges or moldings to provide soft, even illumination of the ceiling above.

Strip lights are also available for 12VDC power, to create lighting effects in customized automobiles and trucks. These strip lights are available in various colors in addition to white. Many have multicolor capability and can be controlled with a handheld remote.

Comparisons

The advantages of an **incandescent lamp** are listed in “[Relative Advantages](#)” on page 179, while advantages of **fluorescent lights** are listed in “[Comparisons](#)” on page 194. These lists can be compared with the following advantages for LED area lighting:

- While the life of an incandescent lamp for room lighting can be as little as 1,000 hours, LED area lighting typically claims up to 50,000 hours.
- The lifetime of an incandescent bulb is the average time it can emit light before catastrophic failure. The lifetime of an LED is the average time it can emit light before gradually dimming to 70% of its rated output. This is a much gentler, less inconvenient failure mode that does not require immediate replacement.
- Unlike a fluorescent light or incandescent bulb, the LED does not contain hot tungsten that fails as a result of erosion.

- Unlike a fluorescent light, an LED does not contain mercury, and therefore does not require special recycling arrangements that entail associated fees.
- While fluorescents can have difficulty starting in low temperatures, an LED is not sensitive to a cold environment.
- Bright LEDs are available in a wide range of colors that do not require filtering. Filters greatly reduce the efficiency of incandescent bulbs when they are used in applications such as traffic signals or rear lights on automobiles.
- High-brightness LEDs can be dimmable. Fluorescent lights are usually not dimmable, or perform poorly in this role.
- LEDs are inherently directional, because the die radiates light at an angle of 90 degrees to its plane. This makes it ideal for ceiling mounting, where as much light as possible should be directed downward. A fluorescent tube or incandescent bulb often requires a reflector which reduces the overall efficiency.
- LEDs are insensitive to cycling. The life expectancy of an incandescent bulb or (especially) a fluorescent tube is reduced by cycling it on and off.
- No flickering. Fluorescent tubes may start to flicker as they age.
- No electrical interference. Fluorescent tubes can interfere with AM radio reception and some audio devices.
- Safe from breakage. LED area lighting does not necessarily use any glass.

However, high-brightness LEDs still have some barriers to overcome:

- Cost. In the United States, before 60W incandescent bulbs were legislated out of existence, they could be sold profitably for less than \$1 each. A T8 fluorescent tube, measuring 1" diameter and 48" long, currently

costs between \$5 and \$6 (retail) but has a life expectancy in the region of 25,000 hours, and uses only 20% of the power of an incandescent bulb to generate two to three times as much light. Clearly the fluorescent tube is a more economical choice, despite the price of the electronics that must be included in the fixture to start the tube. By comparison, currently the purchase price of an LED tube is three times that of a fluorescent tube. It may last twice as long, but is not significantly more efficient, generating perhaps 100 lumens per watt while a fluorescent is typically capable of 90 lumens per watt. Prototype high-brightness LEDs have exceeded 200 lumens per watt, and should be competitive with fluorescents by 2020, but even then, migration will take time.

- Heat sensitivity. Heat reduces the light output and the lifespan of LED fixtures.
- Placement issues. Because LEDs are heat sensitive, they must be installed in locations that do not become excessively hot, their heat sinks must be correctly oriented, and they must have adequate ventilation.
- Color shift. Heat and age may cause the color temperature of an LED to shift slightly, as the color is usually derived from two types of phosphors.
- Nonuniformity. Manufacturing inconsistencies can cause LEDs of the same type to display slightly different color temperatures. Fluorescents and incandescents are more uniform.
- Lower heat output than incandescents. While this is an advantage from the point of view of efficiency, it can be a disadvantage in applications such as traffic signals or airport runway lighting where waste heat can help to keep the lights free from snow or ice.

Values

Although the output from an LED area light is directional, while the output from an incandescent bulb or a fluorescent light is omnidirectional, the intensity is measured the same way in each instance, using *lumens*. This unit expresses the total light emission, without taking directionality into account. (The intensity of **LED indicators** is calibrated in candelas, which measure the power within an angle of dispersion; but candelas are not used for area lighting.)

Typical values for incandescent bulbs are 450 lumens for a power consumption of 40 watts, 800 lumens for a consumption of 60 watts, 1,100 lumens for a consumption of 75 watts, and 1,600 lumens for a consumption of 100 watts. Because much of the output from an incandescent bulb may be wasted by using inefficient reflectors or allowing the light to shine in directions where it is not needed, a high-brightness LED rated at 1,000 lumens may actually appear brighter than a 75-watt incandescent bulb.

A T8 fluorescent tube measuring 48" long by 1" in diameter consumes only 32 watts but emits almost 3,000 lumens—when it is new. This output gradually diminishes by as much as 40% over the lifetime of the tube.

Incandescent bulbs deliver between 10 and 15 lumens per watt, approximately. A new fluorescent tube produces around 80 to 90 lumens per watt, and LED area lighting at the time of writing can provide 100 lumens per watt, under real-world conditions.

What Can Go Wrong

Wrong Voltage

Many high-brightness LED lighting units can be used with either 115VAC or 230VAC. There are exceptions, however. Check the specifications to make sure. Also, it is important to avoid applying domestic supply voltage to 12V LED miniature reflector-bulbs that are intended to replace 12VAC halogen bulbs of the same size.

Overheating

If a high-brightness LED fixture is equipped with a heat sink, this must be exposed to freely flowing air. Any vanes on the heat sink should be oriented vertically to encourage convection, and the fixture must not be placed in an enclosure. Overheating will radically shorten LED life.

Fluorescent Ballast Issues

A fluorescent fixture contains a *ballast* to limit the tendency of the tube to draw excessive current. The ballast is contained in a plastic box attached to the back of the frame in which the tube is mounted.

A *magnetic ballast* contains a coil, and is bypassed by an additional *starter* that applies unlimited current for one second when the power is switched on, preheating the tube to initiate plasma discharge.

An *electronic ballast* performs the same function without a separate starter.

Some LED tubes designed as substitutes for fluorescent tubes may allow a magnetic ballast to remain in the circuit, but may not tolerate an electronic ballast. Other LED tubes require any type of ballast to be unwired from the circuit. The unwiring operation will require disconnection of a couple of wires by removing wire nuts (assuming that the fixture has been designed to comply with U.S. building codes). The wires are then reconnected to apply power directly to the tube, and the wire nuts are reapplied to complete the new connection. The ballast can remain passively in the fixture.

Failing to remove the ballast and/or the starter from a fluorescent fixture before installing an LED tube that requires direct connection to the power supply can damage the tube. Connecting the power incorrectly to the LED tube may result in it failing to light up. Documentation supplied with the LED tube should provide guidance for disconnecting the ballast and connecting the tube. Note that the pin functions on LED tubes are not standardized at this time.

Misleading Color Representation

Because the spectrum of a white LED is not evenly weighted across all wavelengths, it will fail to represent some colors accurately, as shown pre-

viously. This can be important if LEDs are used to illuminate full-color printing or artwork, or if they are installed in stores selling merchandise such as clothes, furnishings, or food.

LED display

In this encyclopedia, a component consisting of multiple separately discernible light-emitting diodes, such as a seven-segment numeral, 14- or 16-segment alphanumeric character, a dot-matrix character, or a display module containing multiple characters, is categorized as an **LED display**. The term *light-emitting diode* is hardly ever used to describe an LED display, as the LED acronym has become ubiquitous. The acronym does not usually include periods between the letters.

An **LED indicator** is defined here as a component usually 5mm or smaller in diameter, made of transparent or translucent epoxy or silicone, most often containing one *light-emitting diode*. It is purposed as a status indicator in a device, rather than as a source of illumination, and is sometimes referred to as a *standard LED*.

LEDs that are designed to illuminate large living or working areas are discussed in a separate entry as **LED area lighting**. They are sometimes referred to as *high-brightness LEDs* and almost always emit white light.

The term **OLED** is an acronym for *Organic Light-Emitting Diode*, a thin panel in which an organic compound is contained between two flat electrodes. Despite its functionality as a form of LED, its design is similar to that of thin-film electroluminescent light sources. Therefore it is discussed in the entry on **electroluminescence**.

OTHER RELATED COMPONENTS

- **LED indicator** (see [Chapter 22](#))
- **LED area lighting** (see [Chapter 23](#))
- **vacuum-fluorescent** (see [Chapter 25](#))
- **electroluminescence** (see [Chapter 26](#))
- **LCD** (see [Chapter 17](#))

What It Does

An LED display presents information on a panel or screen by using multiple segments that emit light in response to a DC current, almost always at a voltage ranging between 2VDC and 5VDC. The display may contain alphanumeric characters and/or symbols; simple geometrical shapes; dots; or pixels that constitute a *bitmap*.

A *liquid-crystal display*, or **LCD**, serves the same purpose as an LED display and may appear very

similar, except that a liquid crystal reflects incident light while an LED emits light. The increasing use of backlighting with LCDs has made them appear more similar to LED displays.

There is no schematic symbol to represent an LED display. Where a segmented display is used, often the segments are represented with drawn outlines.

The simplest, most basic, and probably the best-known example of an LED display is the *seven-*

segment numeral, one of which is shown in [Figure 24-1](#). This is a Kingbright HDSP-313E with a character height of 0.4”.

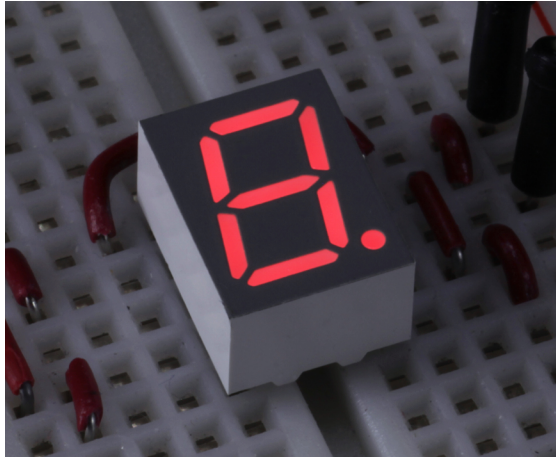


Figure 24-1. The most basic LED display, able to create numerals from 0 through 9 using seven light-emitting segments that can be illuminated individually. An eighth segment forms the decimal point.

How It Works

The process by which an LED generates light is explained in [“How It Works” on page 207](#), in the entry dealing with **LED indicators**. Each light-emitting diode in an LED display is functionally the same as the diode in an LED indicator.

LEDs must be driven with DC. This is a primary distinction between an LED display and an LCD, which requires AC.

Variants

LCD comparisons

LCDs and LED displays can look very similar. This raises the obvious question: which is appropriate for a particular application?

LCDs (without backlighting) are more appropriate for applications such as digital watches and solar-powered calculators where power consumption must be minimized. They are capable of running for years from a single button cell.

LCDs are easily visible in bright ambient light, where LED displays are not. LCDs can also be designed to display complex pictographic shapes and symbols, while the segments of an LED display are more constrained to be simple in shape.

An LCD is more likely to be affected by temperature than an LED, and powering it entails some slight inconvenience, because it requires an AC source that is unlikely to be useful elsewhere in a circuit. If the LCD uses LED backlighting, it will also require a low-voltage DC power source for the backlight. An LED display is easier to use in that it can be driven directly from a microcontroller or logic chip, with only some series resistors to limit the current, and the addition of transistors to provide additional power where necessary.

Seven-Segment Displays

Early seven-segment LED displays were used in digital calculators, before LCDs became an affordable, practical alternative that greatly extended battery life. Initially, the size of the diodes was limited, sometimes requiring magnifying lenses to make them legible.

Seven-segment displays are still used in some low-cost applications, although LCDs have become more common.

[Figure 24-2](#) shows how the segments are identified with letters *a* through *g*. This scheme is used universally in datasheets, and is also used for LCDs. The decimal point, customarily referred to as “dp,” is omitted from some displays. The segments are slanted forward to enable more acceptable reproduction of the diagonal stroke in numeral 7.

Although seven-segment displays are not elegant in appearance, they are functional and are reasonably easy to read. They also enable the representation of hexadecimal numbers using letters A, B, C, D, E, and F (displayed as A, b, c, d, E, F because of the restrictions imposed by the small number of segments), as shown in [Figure 24-3](#).

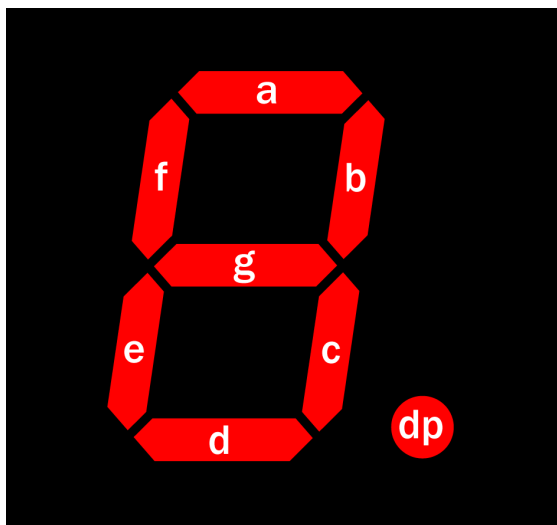


Figure 24-2. A seven-segment LED display. The lower-case identifying letters are universally used in datasheets.

In appliances such as microwave ovens, very basic text messages can be displayed to the user within the limitations of seven-segment displays, as suggested in [Figure 24-4](#).

Numbers 0, 1, and 5 cannot be distinguished from letters O, I, and S, while letters containing diagonal strokes, such as K, M, N, V, W, X, and Z, cannot be displayed at all.

Multiple Numerals

Displays consisting of a single numeral are now rare, as few applications require only one digit. Displays of two, three, and four digits are more common, as shown in [Figure 24-5](#).

Additional Segments

Displays with 14 or 16 segments were introduced in an effort to enable the representation of all the letters of the alphabet. The segment layout of these LED displays is identical to that of comparable LCDs. The differences between 14-segment and 16-segment displays are shown in [Figure 24-6](#). Some are angled forward like seven-segment displays, even though the addition of diagonal segments makes this unnecessary for display of characters such as numeral 7.

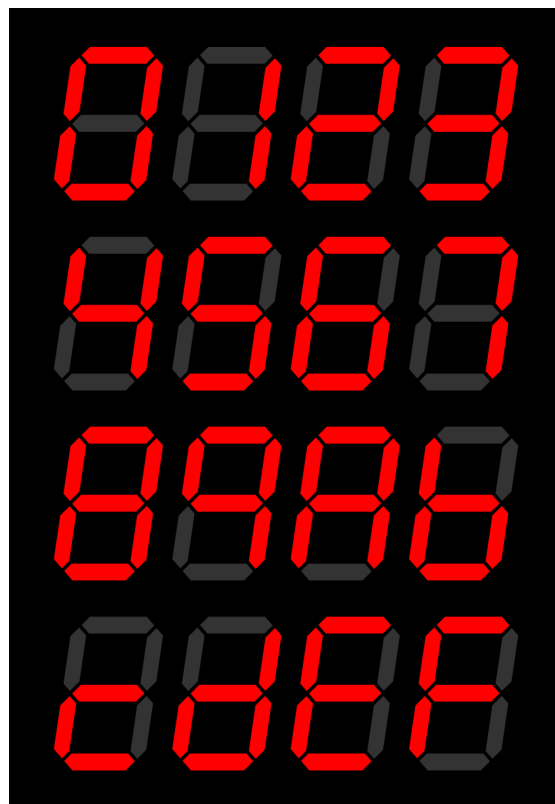


Figure 24-3. Numerals and the first six letters of the alphabet created with seven-segment displays.

[Figure 24-7](#) shows the scheme for identifying the segments of a 16-segment display. This naming convention is used in all datasheets. The lower-case letters that were customary with seven-segment displays are usually abandoned in favor of uppercase, perhaps to avoid confusion with the letter L. Note that letter I is omitted from the sequence.

For a complete alphanumeric character set enabled by a 16-segment display, see [Figure 17-9](#) in the entry discussing LCDs.

An example of a 16-segment alphanumeric LED display is shown in [Figure 24-8](#), mounted on a breadboard and wired to show the letter N. This is a Lumex LDS-F8002RI with a character height of 0.8". The component is still available at the time of writing, but in limited quantities.

Generally speaking, 16-segment displays were never very popular, because the gaps between adjacent segments impaired legibility. LED versions remain more readily available than LCD versions, but dot-matrix displays allow a better-looking, more easily legible alphabet, with the added possibility of simple graphics.

Dot-Matrix Displays

In the 1980s, some personal computers used a video character set in which each letter, numeral, punctuation mark, and special character was formed on a video screen from a fixed-size matrix of dots. A similar alphabet is now used in LED dot-matrix displays (and LCDs, as shown in Figure 17-10).



Figure 24-4. Basic text messages can be generated with seven-segment displays, although they cannot represent alphabetical letters containing diagonal strokes.

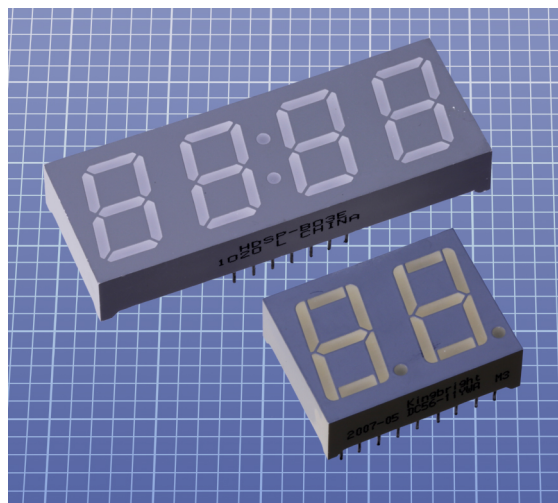


Figure 24-5. Multiple seven-segment LED displays are often combined in a single component. Top: An Avago 2.05VDC 20mA display designed for a clock. Bottom: A Kingbright two-digit display which draws 20mA at 2.1VDC. The unlit outlines of the numerals would normally be hidden behind panels that are tinted to the same colors emitted by the LED segments when lit.

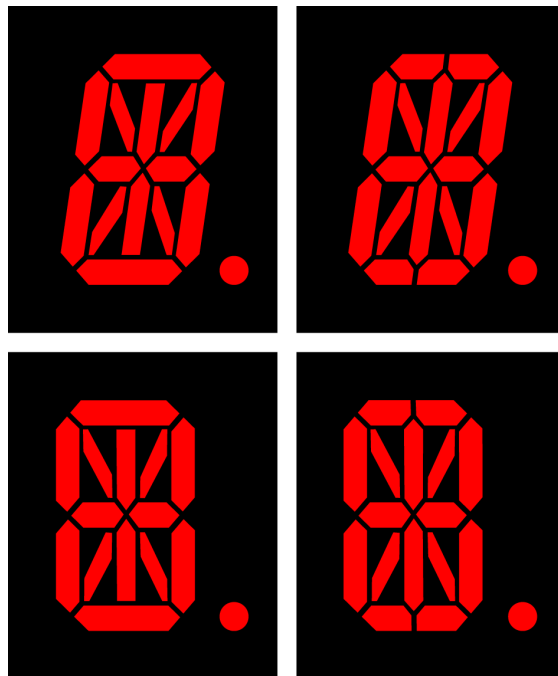


Figure 24-6. Layouts for 14-segment and 16-segment alphanumeric LEDs are identical to those of LCDs.

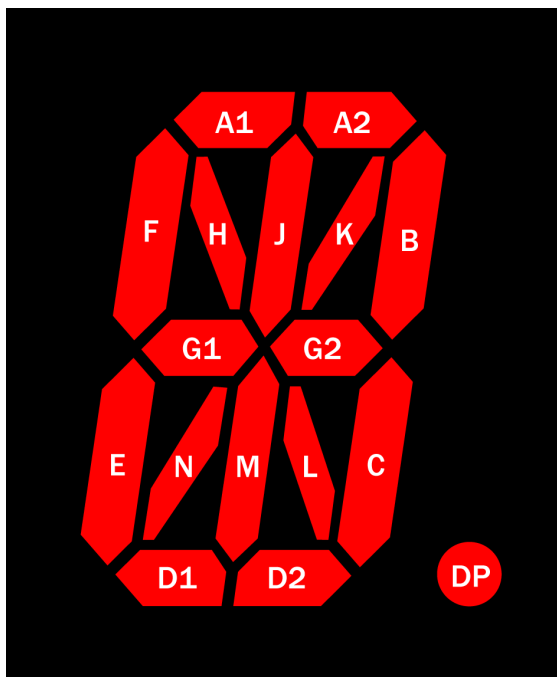


Figure 24-7. The scheme for identification of segments in a 16-segment alphanumeric LED display.

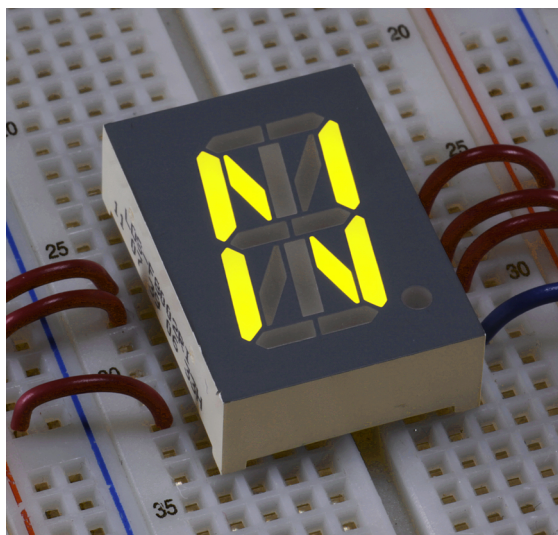


Figure 24-8. A 16-segment alphanumeric LED display showing the letter N.

Alphanumeric dot-matrix characters are often grouped in two or more rows with eight or more

characters per row. The number of characters is always listed before the number of rows, so that an 8x2 display would contain eight alphanumeric characters in two horizontal rows. This type of component is properly described as a *display module*.

Display modules are used in consumer electronics products such as a stereo receiver where simple status messages and prompts are necessary—for example, to show the tone control settings or the frequency of a radio station. Because the cost of small, full-color, high-resolution LCD screens has been driven down rapidly by the mass production of cellular phones, and because these high-resolution screens are much more versatile, they have already displaced dot-matrix display modules in many automobiles and are likely to follow a similar path in other devices.

Pixel Arrays

The 8x8 pixel array of LED dots shown in [Figure 24-9](#) measures 60mm square (slightly more than 2") and contains 64 LEDs, each approximately 5mm in diameter. Similar arrays are available in other sizes and with different numbers of dots. Displays of the same type may be assembled edge-to-edge to enable scrolling text or simple graphics.

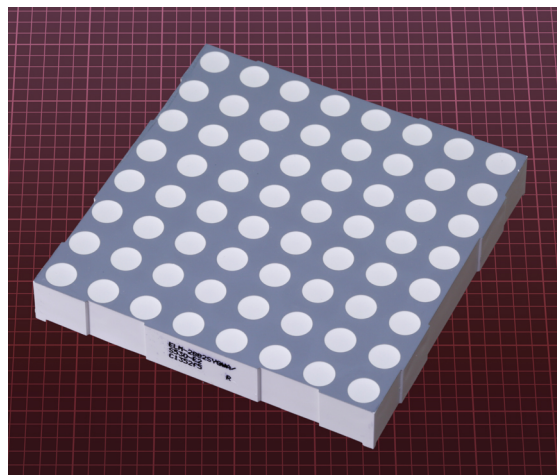


Figure 24-9. An 8x8 matrix of LED dots measuring more than two inches square.

Multiple Bar Display

A bar display is a row of small rectangular LEDs in a single component. It may be used for digital representation of an analog signal. The higher the voltage of the signal, the more bars will be illuminated. A typical application would be to show the signal strength of an input to an audio recorder. Ten bars are often used, as in the display shown in [Figure 24-10](#), but multiple components can be combined end to end.

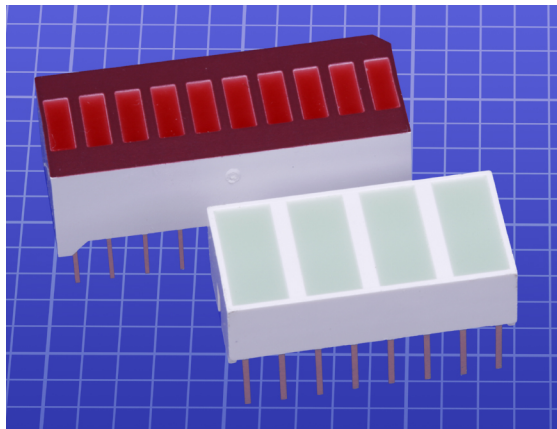


Figure 24-10. Two LED bar displays in which segments can be lit individually.

Single Light Bar

A light bar can be thought of as a single-source LED, as it is configured as a single square or rectangle. It is mentioned here, rather than in the entry for single-source LED indicators, because variants may be subdivided into two, three, four, or (sometimes) more discrete sections. These variants are often included in the same datasheet as the monolithic version.

A light bar contains multiple LEDs (often, four in number) behind a translucent panel that provides evenly diffused radiance.

Values

The values for most LED displays are basically the same as for LED indicators, in terms of color,

brightness, current consumption, and voltage. See [“Values” on page 211](#) for information.

Multiple-character dot-matrix LED display modules may have different requirements for forward voltage and forward current, depending on drivers that are incorporated in the module. Because there is no standardization for these modules, it will be necessary to consult the manufacturer’s datasheet.

How to Use It

Seven-Segment Basics

The diodes in a seven-segment LED display share either a [common anode](#) or a [common cathode](#), the latter being more frequently used. The two types of internal wiring are provided for convenience only. Externally, the displays function identically.

A schematic suggesting the internal wiring and pinouts of a typical ten-pin common-cathode display is shown in [Figure 24-11](#). The pins are numbered as seen from above. Appended to each number is the identity of the segment to which it is connected. Pins 3 and 8 are connected with the cathodes of all the internal LEDs. Both of these pins should be used, to serve as heat sinks for the display.

Note that series resistors are not included inside the display and must be added externally. Their value will be determined by the power supply, to limit the forward current and forward voltage through the LEDs to the extent specified by the manufacturer.

An encapsulated [resistor array](#) containing either seven or eight resistors in an SIP or DIP chip can be used instead of individual resistors. A seven-segment LED display would require the type of resistor array in which both ends of each resistor are accessible.

Where two or more numerals are combined in a single component, this type of display is likely to have two horizontal lines of pins. In this case, pin 1 will be at the bottom-left corner, seen from

above. As always, the pins are numbered counterclockwise, seen from above.

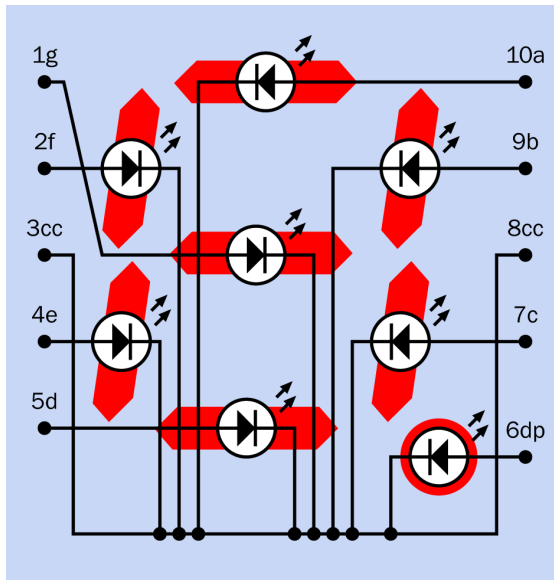


Figure 24-11. A schematic view of internal connections and pinouts of a seven-segment common-cathode LED display. The numbers refer to the pins of the component, seen from above. The 1 pin may also have a mark beside it on the component, for identification. The orientation of the display can be deduced if there is a decimal point, as this should be at the bottom-right corner.

Where three or more numerals are combined in a single component, the pinouts may be designed for multiplexing rather than individual access to every segment of each numeral. A four-digit clock display, for instance, may have seven pins that connect in parallel to respective segments in all of the numerals, and four additional pins that can ground each numeral in turn, so that they can be selected sequentially.

Driver Chips and Multiplexing

Illuminating the appropriate segments in a single numeral can be done directly from a microcontroller, or through a driver chip such as the well-known and widely used 4543B that converts a binary-coded decimal input into appropriate segment output patterns. The chip can source sufficient current to drive each segment through

a series resistor. Its pinouts are shown at [Figure 24-12](#).

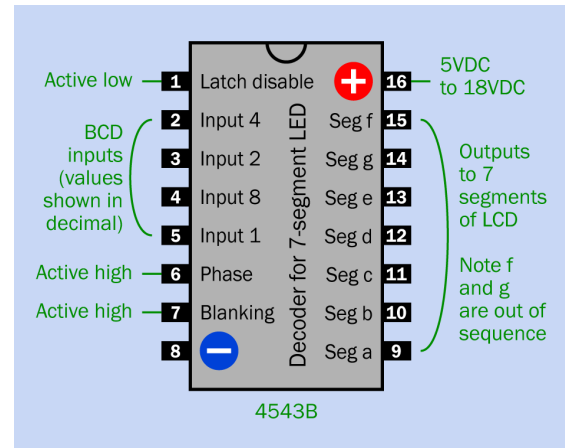


Figure 24-12. Pinouts of the 4543B seven-segment LED driver chip.

When used in conjunction with a microcontroller, the 4543B can drive several seven-segment displays by multiplexing them. The basic schematic to achieve this is shown in [Figure 24-13](#), omitting optional features such as leading-zero blanking or connections for a decimal point. The microcontroller sends the binary code for the first numeral and simultaneously grounds the common cathode of that numeral through a transistor, which is needed because as many as seven segments of the numeral may be passing current in parallel. The microcontroller then sends the binary code for the second numeral, and grounds it; then sends the binary code for the third numeral, and grounds it; and the cycle repeats. So long as this process is performed at sufficient speed (at least 50Hz), persistence of vision will create the illusion that all the numerals are active simultaneously. The circuit can be compared with a similar circuit to drive LCDs, shown in [Figure 17-17](#).

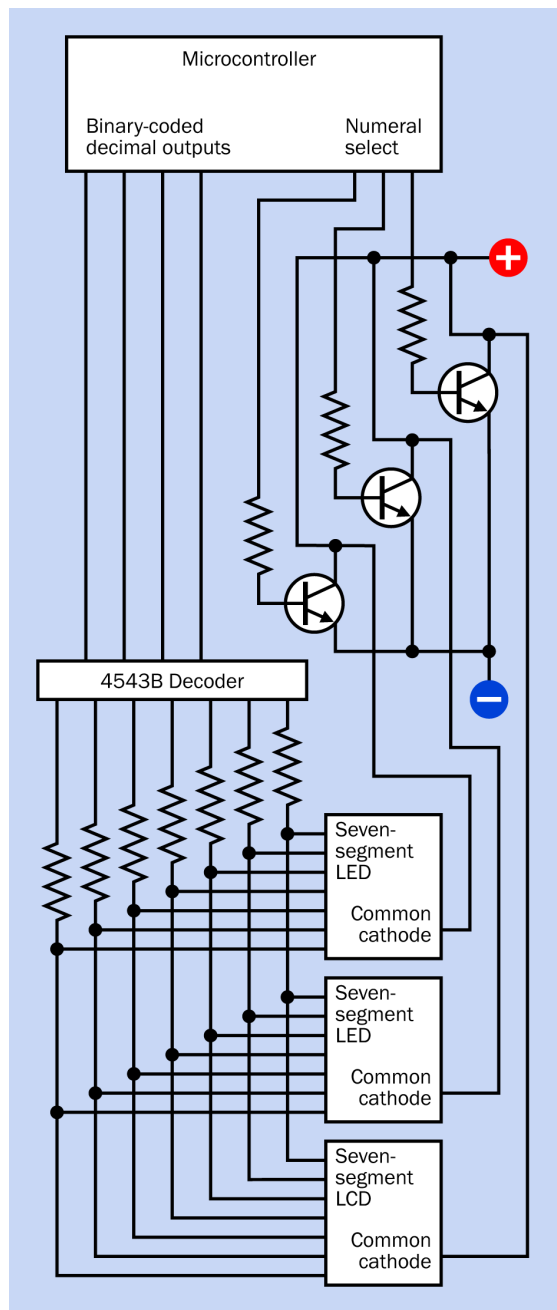


Figure 24-13. A basic, simplified circuit for driving multiple seven-segment LED displays by multiplexing them.

The disadvantage of this system is that the microcontroller must update the numerals constantly while performing other duties. To reduce this burden, a “smarter” driver such as the

MC14489 can be used, controlling up to five 7-segment digits, or the Intersil ICM7218, which can control up to eight 7-segment digits.

The MC14489 controller receives data serially, using SPI protocol, and handles the details of addressing the LEDs. Because it contains latches to sustain the displayed data, a microcontroller only needs to communicate with the driver when the displayed information needs to be updated.

The ICM7218 is a more sophisticated chip, available in several variants, one of which can receive data on an 8-bit bus and run the seven-segment displays in hexadecimal mode.

Sixteen-Segment Driver Chip

The MAX6954 by Maxim can drive up to eight 16-segment alphanumeric LED displays using a scheme known as Charlieplexing, named after a Maxim employee named Charlie Allen who came up with the concept as a way of reducing the pin count required for multiplexing. Other Maxim controllers use this same protocol, which is transparent to the user.

A microcontroller sends data serially via I2C protocol to the MAX6954, which contains a variety of features. It can drive 14-segment and 7-segment displays as well as 16-segment displays, and contains a 104-character alphabet for each of them. Setting up a microcontroller to send the various necessary command codes to the MAX6954 is not a trivial matter, and bearing in mind the probably impending end-of-life of 16-segment displays, a better option may be to use dot-matrix LED display modules that have controller logic built in.

Dot-Matrix LED Display Modules

A dot-matrix LED display module requires data to define a character set, and a command interpreter to process instructions that will be embedded in a serial data stream. These capabilities are provided either by separate chips or (more often) are incorporated into the LED display module itself.

The SSD1306 is a monochrome graphical controller capable of I2C or SPI serial communication, or parallel communication. When this capability is built into a display module, only one of these types of communication may be activated.

The SSD1331 is a color graphical controller with similar communication capabilities.

The WS0010 is a monochrome controller, compatible with HD44780, which is designed to control LCDs.

Typical controller functions are summarized in “[Alphanumeric Display Module](#)” on page 168. Because there is no standardization in this field, precise details must be found in manufacturers’ datasheets.

Pixel Arrays

The connections inside an 8x8 pixel array are shown in [Figure 24-14](#), where the schematic symbols for LEDs have been replaced by gray circles for space reasons. To illuminate one LED, power is supplied to the intersection where it resides. In the figure, each vertical conductor (identified as A1, A2 . . . A8) can power the anodes of a column of eight LEDs, while each horizontal conductor (identified as C1, C2 . . . C8) can ground the cathodes of a row of eight LEDs. If only one vertical conductor is connected with positive power while one horizontal conductor is grounded, only one LED will light up, at the intersection of the active conductors.

A problem occurs if we wish to illuminate two LEDs. Suppose they are located at (A3,C2) and (A6,C5). Unfortunately, providing power to them will also result in activating LEDs at (A3,C5) and (A6,C2), as shown in [Figure 24-15](#), where the yellow circles represent LEDs that have been switched on.

The answer to this problem is to rasterize the process. In other words, data is supplied on the array one line at a time, as in the process by which a TV picture is generated. If this is done quickly enough, persistence of vision will create the

illusion that the LEDs are illuminated simultaneously.

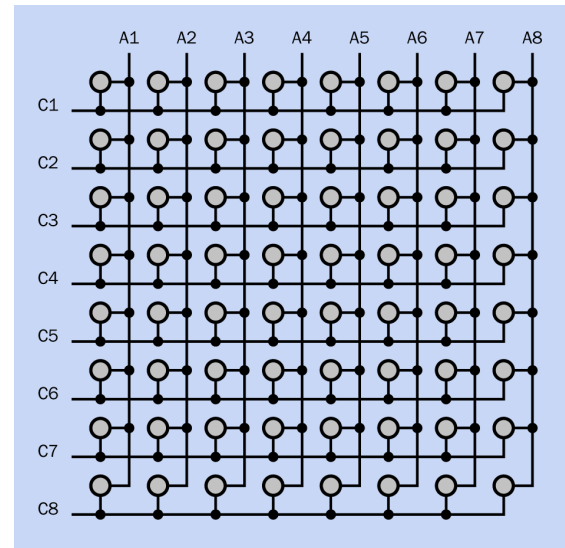


Figure 24-14. Internal connections in the 8x8 matrix. Each gray circle represents an LED.

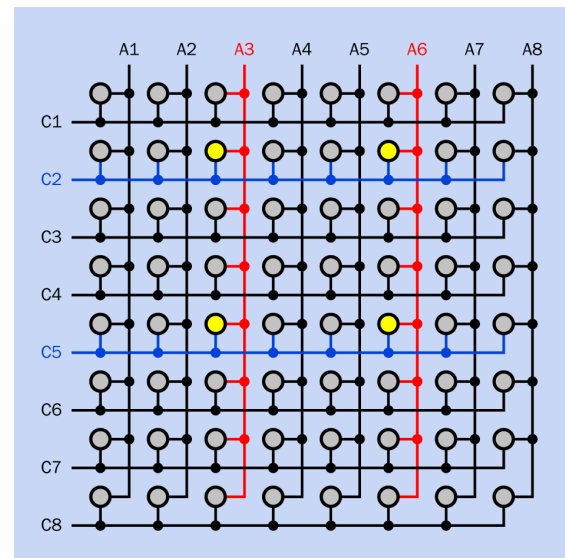


Figure 24-15. An attempt to illuminate LEDs at (A3,C2) and (A6,C5) will also activate the LEDs at (A3,C5) and (A6,C2).

A form of multiplexing is used to achieve this. One row of LEDs is connected to negative ground

for a brief interval. During this interval, the anodes of selected LEDs are powered momentarily. Then the next row is grounded, and selected LEDs along that row are powered momentarily. This process is repeated for all eight rows before being repeated.

If several 8x8 matrices are assembled edge to edge, their horizontal conductors can be common to all of them. A horizontally scrollable display (sometimes referred to by the archaic term, *electric newspaper*) would then be possible, although the circuit design would be nontrivial.

Multiple Bar Display Driver

The LM3914 is a driver for a bar display that compares an analog input with a reference voltage and provides power to the segments of a multiple bar display, ranging from 2mA to 30mA, adjustable to match the specification of the display that is being used. The chip can generate either a “thermometer” effect, as more outputs are activated when the analog input increases, or a “moving dot” effect, in which only one output is on at a time.

One-Digit Hexadecimal Dot Matrix

While multi-character dot-matrix LED display modules are a versatile way to display prompts and numbers, a simpler component is sometimes sufficient. The Texas Instruments TIL311 is a minimal dot-matrix LED display that receives a binary value from 0000 through 1111 on its four input pins and generates the output in hexadecimal form, using numerals 0 through 9 and letters A through F. The sixteen possible outputs in the dot-matrix display are shown in [Figure 24-16](#). Although this component is no longer being manufactured, it is widely available from many sources, especially in Asia. It eliminates the series resistors and controller chip that are customary for a seven-segment display, and has a better-looking output.



Figure 24-16. Sixteen possible outputs that can be displayed by the Texas Instruments TIL311 to show a hexadecimal value in response to a four-digit binary input.

A sample of the TIL311 is shown in [Figure 24-17](#) displaying the number 2.

If two or more of these chips are put together, they can be multiplexed to display multi-digit decimal or hexadecimal integers.

The chip features two decimal points, one to the left of the displayed numeral, and one to the right. If they are activated, they require their own series resistors to limit the current.

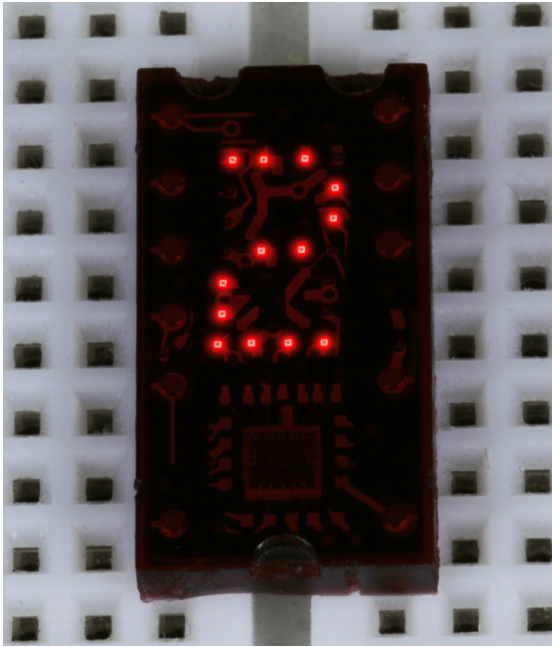


Figure 24-17. The Texas Instruments TIL311 can be driven directly by a microcontroller or counter chip, with no series resistors necessary. It can generate a hexadecimal output.

What Can Go Wrong

Common Anode versus Common Cathode

An LED display containing a common cathode is usually identical in appearance to a display containing a common anode, and the two versions will be distinguished by only one digit or letter in their part numbers. Because LED displays have a limited tolerance for reverse voltage, part numbers should be double-checked before applying power.

Incorrect Series Resistance

A common error is to assume that only one series resistor is necessary for a seven-segment LED

display, either between the common cathode pin and ground, or, if there is a common anode, between that and the positive power supply. The problem is that if the resistor is suitable for a single LED, its value will be too high when several segments of the display are sinking current or drawing current through it. If its value is reduced, it will be too low when only two segments are using it (as when generating the number 1).

To provide equal illumination of all the segments, each must have its own series resistor.

Multiplexing Issues

When several displays are multiplexed, they naturally appear dimmer, creating a temptation to compensate by upping the current. Because current is only being applied to each display intermittently, a natural assumption is that a higher current can be safely used.

This may or may not be true. When running an LED device with pulsed current, the peak junction temperature, not the average junction temperature, determines the performance. At refresh rates below 1kHz, the peak junction temperature is higher than the average junction temperature, and the average current must therefore be reduced.

Datasheets must be checked to determine whether a device is designed with multiplexing in mind and, if so, what the recommended peak current is. Very often this value will be accompanied by a maximum duration in milliseconds, and a calculation may be necessary to determine the refresh rate, bearing in mind how many other LED displays are being multiplexed in the same circuit at the same time.

Irresponsible multiplexing will shorten the life of an LED display or burn it out.

vacuum-fluorescent display

The term **vacuum-fluorescent display** is seldom hyphenated, but the first two words are hyphenated here as they constitute an adjectival phrase. The acronym *VFD* is becoming increasingly popular, although it is ambiguous, being also used to identify a *variable frequency drive*. In both instances, the acronym is printed without periods between the letters.

The entry in this encyclopedia dealing with **fluorescent lights** does not include VFDs, because their purpose and design are very different. A VFD is an informational display, often showing numerals and letters, while a fluorescent light merely illuminates a room or work area. Although a VFD does use fluorescent phosphors, they are printed onto light-emitting segments of the display instead of being applied to the inside surfaces of a glass envelope.

OTHER RELATED COMPONENTS

- **LED indicator** (see [Chapter 22](#))
- **LCD display** (see [Chapter 17](#))
- **electroluminescence** (see [Chapter 26](#))

What It Does

A **vacuum-fluorescent display** or *VFD* superficially resembles a backlit monochrome **LCD** or an **LED display**, as it can represent alphanumeric characters by using *segments* or a *dot matrix*, and can also display simple shapes. It is often brighter than the other information display systems, and can emit an intense green phosphorescent glow that some people find aesthetically pleasing, even though a grid of very fine wires is superimposed internally over the displayed image.

There is no specific schematic symbol to represent a vacuum-fluorescent display.

How It Works

The display is mounted inside a sealed capsule containing a high vacuum. A widely spaced series of very fine wires, primarily made of tungsten, functions as a *cathode*, moderately heated to encourage electron emission. The wires are often referred to as *filaments*.

A **fluorescent light** uses AC, and both of its electrodes are often confusingly referred to as cathodes. A VFD uses DC, and its cathode array has the function that one would expect, being connected with the negative side of the DC power supply.

Opposite the cathode, just a few millimeters away, is an anode that is subdivided into visible alphanumeric segments, symbols, or dots in a matrix. Each segment of the anode is coated with *phosphors*, and individual segments can be sep-

arately energized via a substrate. When electrons strike a positively charged anode segment, it emits visible light in a process of *fluorescence*. This behavior can be compared with that of a *cathode-ray tube*. However, the cathodes in a VFD are efficient electron emitters at a relatively low temperature, while the cathodes in a cathode-ray tube require substantial heaters.

Anode, Cathode, and Grid

A *grid* consisting of a mesh of very fine wires is mounted in the thin gap between the filaments of the cathode and the segments of the anode. A simplified view of this arrangement is shown in *Figure 25-1*.

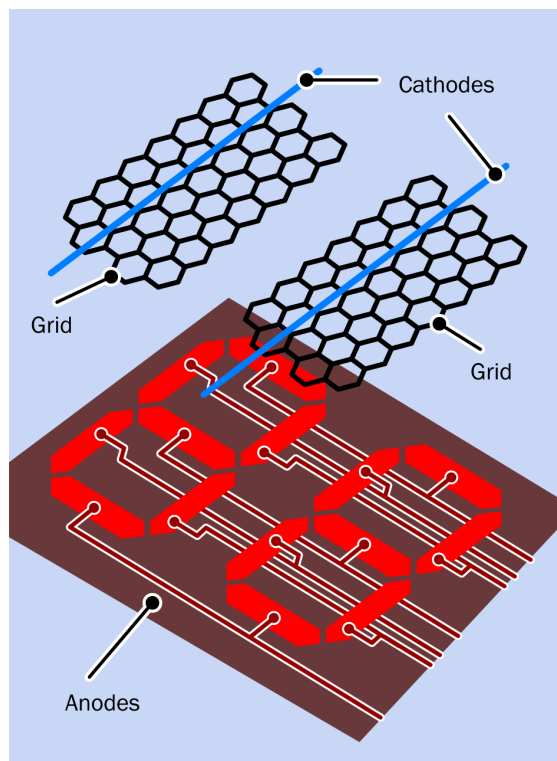


Figure 25-1. The basic elements of a vacuum-fluorescent display.

The polarity of the charge on the grid controls and diffuses electrons emitted by the cathode. If a grid section is negatively charged, it repels electrons and prevents them from reaching the

sections of the anode beneath it. If the grid section is positively charged, it encourages electrons to reach the anode. Thus, the grid functions in the same way as the grid in a triode vacuum tube, but its conductors are so thin, they are barely visible.

How to Use It

Electronic calculators used vacuum-fluorescent displays during the 1970s, before LED displays and LCDs became more competitive. Purely numeric VFD modules are still available as strings of digits, although they are becoming uncommon and have been replaced by alphanumeric dot-matrix modules where each VFD character is mounted in its own glass module on a separate substrate.

Figure 25-2 shows the interior of a Commodore calculator from the 1970s, with its nine-digit vacuum-fluorescent display enclosed in one glass capsule.

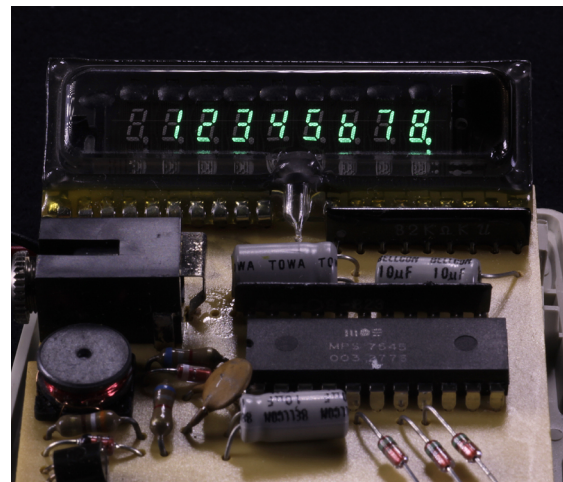


Figure 25-2. The vacuum-fluorescent display from a 1970s Commodore calculator.

A closeup of three digits from the previous figure appears in *Figure 25-3*, showing the grid superimposed above each numeral.



Figure 25-3. Three digits from the previous figure, showing the grid that controls their illumination.

A tinted filter of the same color as the display is usually placed in front of it, to conceal some of its workings. Thus, the Commodore calculator used a green filter in front of its green numerals. [Figure 25-4](#) shows two seven-segment numerals from another device, with the filter removed. This reveals not only the grid but also the horizontal wires that function as the cathode. Connections between the segments of the numerals and a backplane are also visible.

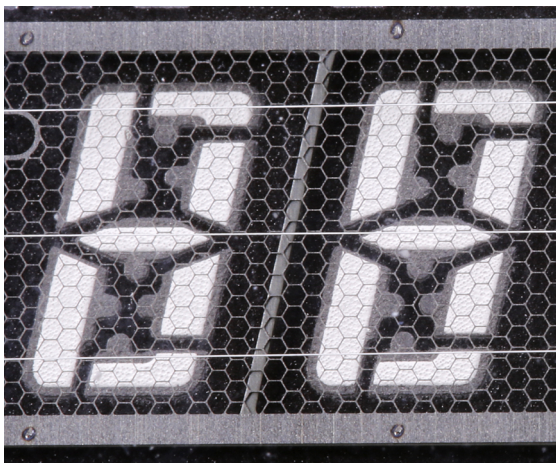


Figure 25-4. Seven-segment numerals viewed without a colored filter, revealing the cathode (horizontal wires) and the grid (wire mesh).

Modern Application

A modern VFD module is likely to be mated with a driver that converts 5VDC to the higher voltage (typically 50VDC to 60VDC) required for the display. Built-in logic may offer the option to receive

data via an 8-bit parallel bus or with SPI serial protocol, and will contain a character set. A typical display resolution is 128 x 64 pixels.

The combination of a grid and a segmented anode enables a VFD to be controlled by [multi-plexing](#). For instance, in a display of four seven-segment numerals, the same equivalent segments in all four numerals can be connected in parallel while a separate grid covers each numeral. When each grid is positively energized, it selects the corresponding numeral, and the on-off segment patterns appropriate to that numeral are supplied. This procedure is repeated for each numeral in turn. Persistence of vision makes it appear that they are all active simultaneously.

Variants

Color

Although a VFD cannot provide a full-color display, selected anode segments can be coated with different phosphor colors, which can fluoresce simultaneously. Two or three individual colors are typically used, as in the display for a CD player where color helps to distinguish a variety of different functions. A closeup of a portion of the display from a CD player (with color filter removed) appears in [Figure 25-5](#).



Figure 25-5. The lefthand section of a vacuum-fluorescent display from a CD player.

Character Sets and Pictorial Design

In the past, VFDs have combined seven-segment numerals in the same display as custom-shaped anodes. Solid-state gain meters in an audio amplifier, for instance, have used numerals beside a pictorial representation of gain levels resembling analog meters. The look and layout of a display of this type has been unique to a particular product.

Modern VFDs tend to use a generic dot-matrix display in which a character set in firmware dictates how patterns of dots are grouped to form numbers, letters, symbols, or icons.

The appearance of character sets generated with generic segments and dot-matrix arrays is thoroughly discussed and illustrated in the entry describing liquid-crystal displays in [Chapter 17](#). VFD alphanumeric modules are identical in visual design to LCD modules, even though the internal electronics are different.

Comparisons

Two advantages of a VFD are that it functions well at low temperatures (unlike an LCD) and has suf-

ficient brightness and contrast to be usable in sunlight (unlike many LED displays). It can be viewed from almost any angle.

Typical applications have included digital instrumentation in automobiles, information displays in audio and video consumer-electronics equipment, and numerical readouts in vending machines, medical devices, and some digital clocks.

Because a VFD requires a relatively high voltage, has significant power consumption, can show only a limited range of fixed colors, and is more expensive than LED displays or LCDs, its popularity has declined since the end of the 1990s.

What Can Go Wrong

Fading

VFDs gradually fade with age, as a result of reduced electron emission from the electrodes or diminishing performance of the phosphor coatings. Increasing the working voltage can prolong the life of a display.

electroluminescence

The field of **electroluminescent** devices is sometimes referred to as *EL*. The same acronym can also be applied as an adjective to an individual electroluminescent device, as in, for example, “an EL panel.”

An *organic light-emitting diode*, more commonly known by its acronym *OLED*, is included in this entry because it is technically an electroluminescent device and its design concept is similar to that of an electroluminescent panel. Generic LEDs are also technically electroluminescent, but are not commonly described as such, and have their own entries in this encyclopedia under the subject categories **LED indicator**, **LED area lighting**, and **LED display**.

OTHER RELATED COMPONENTS

- **LED indicator** (see [Chapter 22](#))
- **LCD display** (see [Chapter 17](#))
- **fluorescent light** (see [Chapter 20](#))
- **vacuum-fluorescent** display (see [Chapter 25](#))

What It Does

An electroluminescent device configured as a *panel*, *ribbon*, or *rope-light* contains phosphors that emit light in response to a flow of electricity.

Panels can be used as backlights for LCD displays or, more often, as always-on low-power devices such as exit signs and night lights. Ribbons and rope lights (the latter being also known, more accurately, as *light wires*) are used mainly as recreational novelties. They can be battery powered through a suitable voltage converter. A battery-powered rope light can be wearable.

Thin-film OLED electroluminescent panels are used in small video screens in handheld devices. At the time of writing, OLED TV screens measuring 50” or more have been demonstrated, but are not yet economic for mass production.

No specific schematic symbol exists to represent any electroluminescent device or component.

How It Works

Luminescence is the emission of light as a result of a process that does not require heat. (The opposite phenomenon is *incandescence*, in which heating causes an object to emit light; see [Chapter 18](#) for a description of **incandescent lamps**.)

Electroluminescence is luminescence resulting from stimulation by electricity. This very broad definition really includes devices such as **LEDs**, although they are hardly ever described in those terms. Electroluminescence generally refers to panels, films, or wires where electrodes are in direct contact with light emitters such as phosphors.

The exception is an *organic LED*, usually known by its acronym *OLED*, which is frequently de-

scribed as an electroluminescent device, perhaps because its configuration as a sandwich of thin, flat layers resembles an electroluminescent panel. Two of the layers are semiconductors, and they interact as light-emitting diodes.

Phosphors

A *phosphor* is a compound such as zinc sulfide that will emit light when it receives an energy input from another light source or from electricity. Typically the compound must be mixed with an *activator* such as copper or silver.

For many decades, TV sets and video monitors were built around *cathode-ray tubes* in which the interior of the screen, at the front of the tube, was coated with phosphors. A beam of electrons that fluctuated in intensity generated a picture on the screen by drawing it as a series of lines.

Derivation

The term *phosphor* is derived from *phosphorescence*, which in turn comes from the name of the element *phosphorous*, which will glow when it oxidizes in moist air. (These terms were established before other forms of luminescence were discovered and understood. The behavior of phosphorous is really an example of *chemiluminescence*.)

For our purposes, a phosphor is a compound that is capable of fluorescence or electroluminescence.

Variants

Panels

Electroluminescent panels using phosphor powder, sometimes referred to as *thick phosphor*, are a popular choice where a constant, uniform, low light output is acceptable.

An electric potential is established between two films that act as electrodes, separated by a layer of phosphor crystals. Some manufacturers refer to this configuration as a *light-emitting capacitor* because the structure resembles a capacitor,

even though that is not its purpose. The front film is transparent, allowing light to escape.

An electroluminescent panel can be powered by AC or DC but requires at least 75V. Its power consumption is self-limiting, so that no control electronics are required other than a voltage converter if battery power is used.

The phosphors generate a constant, evenly distributed luminescence over the entire area, although the output is not very intense. Applications include night-lights, exit signs, and back-lighting for wristwatches.

Panelescent electroluminescent lighting by Sylvania was used for instrument panel displays in some car models such as the Chrysler Saratoga (1960 through 1963) and Dodge Charger (1966 through 1967). It is still used for night-lights. *Indiglo* electroluminescent displays are still widely used in wristwatches.

The interior components of a disassembled electroluminescent night-light are shown in [Figure 26-1](#). The panel emits a natural pale green glow. A separate blue or green filter passes the glow while blocking other colors of incident light that would otherwise reflect off the panel.

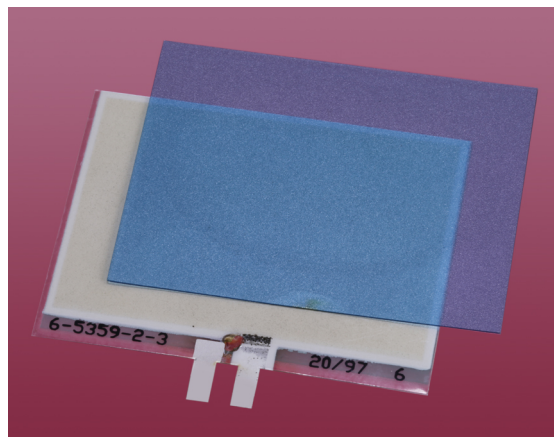


Figure 26-1. The two interior components of an electroluminescent night-light: the luminescent panel, and a separate translucent filter.

Electroluminescent night-lights were popular in the 1970s and 1980s, often featuring cartoon characters to appeal to children. Figures 26-2 and 26-3 show the same night-light in its daytime off-state and its night-time on-state, respectively.



Figure 26-2. A vintage Panelescent brand night-light, several decades old, in its off-state.



Figure 26-3. The same night-light, with its green radiance visible under conditions of low ambient light.

Advantages of electroluminescent panels include the following:

- Low current consumption. One US manufacturer claims that a single exit sign will use electricity costing less than 20 cents per year, while the annual cost of a night-light will be less than 3 cents per year.
- Long life, up to 50,000 hours.

- Self-regulating; no control circuitry required.
- Omnidirectional light output.
- Very wide operational temperature range, between approximately -60 and +90 degrees Celsius.
- Can be plugged directly into a wall outlet.

Disadvantages include:

- Limited light output.
- Very limited choice of colors.
- Not very efficient, 2 to 6 lumens per watt (although the low light output naturally entails low power consumption).
- Gradual reduction in phosphor performance over time.
- High voltage required: 60V to 600V. Ideal for plugging into a wall outlet, but requires a converter when used with battery-powered devices.

Flexible Ribbons

The light-emitting layers inside a night-light are somewhat flexible, and can be made more flexible by reducing their thickness. The result is an electroluminescent ribbon that has some novelty value, and may be used for customizing automobiles. Figure 26-4 shows a ribbon about 1.5" wide and 12" long, designed for 12VDC power applied through an inverter.

Rope Light

A rope light or wire light may resemble a *glowstick*. However, a glowstick generates light from chemiluminescence (chemical reactions that release photons), while a rope light uses electricity.

Figure 26-5 shows a rope light powered by two AA batteries connected through an inverter.

At the center of the rope light is a conductor that serves as one electrode. It is coated in phosphors, and the layer of phosphors is protected by a transparent sheath. One or more thin wires is wrapped around the sheath in a spiral, with large

gaps between one turn and the next. These wires serve as the second electrode. The wires are enclosed in transparent insulation that forms an outer sheath.

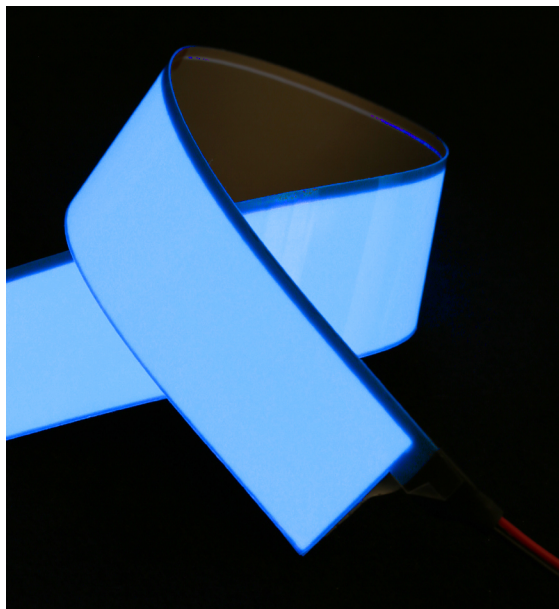


Figure 26-4. A 12" length of electroluminescent ribbon.

When AC is applied between the electrodes, the layer of phosphors emits light that radiates out in the gaps between the thin wires. The color of the light can be modified by using tinted outer insulation.

OLED

An OLED uses two thin, flat electrodes, somewhat like a thick-phosphor electroluminescent panel, except that it contains more layers and is capable of generating more light. The layers in an OLED are “organic” in that they consist of chemically organic molecules containing carbon and hydrogen atoms and generally do not contain heavy metals.

While an LCD video monitor or TV screen must have a separate backlight, an OLED generates its own light. This reduces the thickness of the display to a few millimeters and makes it potentially more efficient.

The semiconductor layers are subdivided into pixels, each functioning as a light-emitting diode, while additional layers carry a matrix of conductors for pixel addressing. In an AMOLED, the conductors form an active matrix, while in a PMOLED, they form a passive matrix.

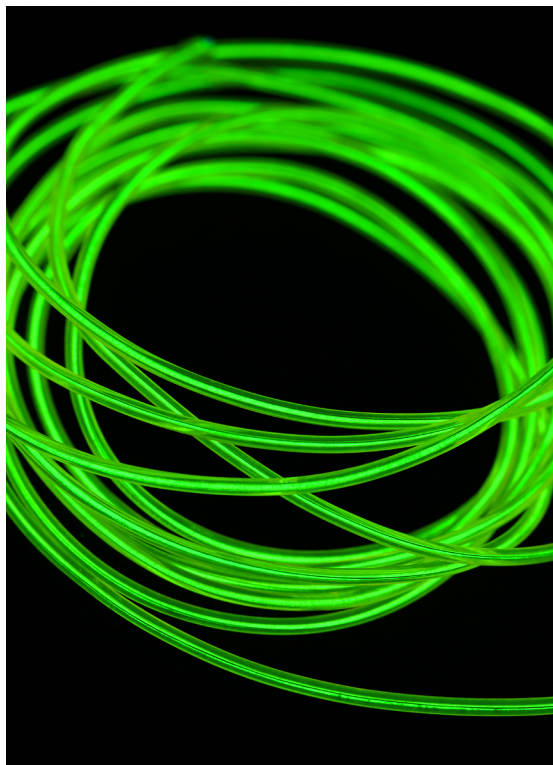


Figure 26-5. A length of glowing rope light, also known as a light wire.

In an active matrix, each pixel is backed with a *thin-film transistor* to store its state while the energizing voltage transitions. This is often described as a *TFT* display; but the term is interchangeable with “active matrix.”

In a passive matrix, each pair of conductors simply supplies current to a pixel. This is cheaper and easier to fabricate but is less responsive.

The terms “active matrix” and “passive matrix” have the same meaning as when used to describe a **liquid-crystal display**.

Monochrome OLED display modules with dot-matrix characters are currently available from China for just a few dollars. Although they appear superficially similar to LCD modules, they generate pure white-on-black characters.

Small full-color OLED screens are used in smartphones and on camera backs, but at the time of writing, large OLED screens are not a mature technology, partly because of production costs. A great variety of chemicals and layer configurations have been tried, and the application of pixels to a substrate has been attempted with vacuum deposition through a shadow mask and with a system similar to inkjet printing. Pixels that emit red, green, and blue light have also been used. Filtered pixels have been used. One dominant process has not yet emerged.

Longevity and brightness have been problems. Where red, green, and blue diodes have been

used, the different colors deteriorate at different rates. While the human eye tolerates an overall reduction in brightness, it does not tolerate a slight color shift caused by blue pixels, for instance, losing brightness more rapidly than red pixels.

Because OLED screens promise to be thinner, lighter, and brighter, and may eliminate the need for a fragile glass substrate, there is a strong incentive to develop this technology, which seems likely to gain dominance in the future.

OLED panels may also become a source of diffuse, shadowless room lighting or office lighting when practical problems have been solved and costs have fallen significantly.

transducer

The term **transducer** is used here to describe a *noise-creating device* that is driven by external electronics. By comparison, an **audio indicator** (discussed in the next entry) contains its own internal electronics and only requires a DC power supply. Either of these components is often described as a *beeper* or *buzzer*.

A **speaker**, more properly termed a *loudspeaker*, is an electromagnetic transducer but is seldom described in those terms. It has a separate entry in this encyclopedia and is defined here as a *sound reproduction device* that is larger and more powerful than a typical transducer and has a more linear frequency response.

While piezoelectric transducers formerly used crystals, only the more modern piezoelectric type that uses a ceramic wafer will be considered here.

Some transducers convert sound into electricity, but these are categorized as *sensors*, and will be discussed in Volume 3. The only transducers discussed in this entry are those that convert electricity into sound.

OTHER RELATED COMPONENTS

- **audio indicator** (see [Chapter 28](#))
- **headphone** (see [Chapter 29](#))
- **speaker** (see [Chapter 30](#))

What It Does

An audio **transducer** is a device that can create an *alert*. It requires an AC signal that is supplied by external electronics, and in its simplest form may be referred to as a *buzzer* or a *beeper*.

Audio alerts are used in microwave ovens, washer/dryers, automobiles, gasoline pumps, security devices, toys, phones, and many other consumer products. They are often used in conjunction with touch pads, to provide audio confirmation that a tactile switch has been pressed.

The schematic symbols in [Figure 27-1](#) can be used to represent any kind of audio alert, including **indicators**, which contain their own electronics to generate a simple tone or series of

tones. Type A is probably the most popular symbol. Types B and C often appear with the word “buzzer” printed beside them for clarification. D and E are really symbols for a **speaker**, but are often used for an alert. F is the symbol for a *crystal*, now sometimes used to indicate a piezoelectric noise maker. G specifically represents an electromagnetic transducer, but is seldom used.

How It Works

A circular diaphragm is glued at its edges inside a cylindrical plastic enclosure, usually measuring from around 0.5” to 1.5” in diameter. The enclosure is sealed at the bottom but has an opening at the top, so that sound can emerge from the upper side of the diaphragm without being par-

tially cancelled by sound of opposite phase that is emitted from the underside of the diaphragm. The enclosure also amplifies the sound by resonating with it, in the same way that the body of a guitar or violin amplifies a note being played on the strings.

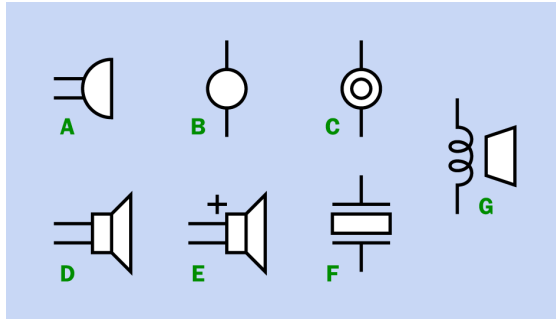


Figure 27-1. An assortment of symbols which can represent a transducer or an indicator. See text for details.

The diaphragm is activated either electromagnetically or piezoelectrically, as described next.

Externally, a transducer may be indistinguishable from an audio indicator such as the one pictured at [Figure 28-1](#).

Variants

Electromagnetic

An electromagnetic transducer contains a diaphragm that is usually made of plastic. Mounted on it is a smaller *ferromagnetic disc* that responds to the fluctuating field from AC passing through a coil. When the diaphragm vibrates, it creates pressure waves that are perceived by the human ear as sound.

A car horn is a particularly loud form of electromagnetic transducer.

Piezoelectric

A piezoelectric transducer contains a diaphragm consisting of a thin brass disc on which is mounted a *ceramic wafer*. When an AC signal is applied between the piezoelectric wafer and the disc, the disc flexes at that frequency.

The term *piezo* is derived from the Greek *piezein*, which means “to squeeze or press.”

Ultrasonic Transducer

The diaphragm in an *ultrasonic transducer* vibrates at a frequency above the range of human hearing. This component may be electromagnetic, piezoelectric, or crystal-based. Often it is used in conjunction with an ultrasonic receiver as a distance measuring device. The two components can be sold pre-mounted on a breakout board. An output from the board can consist of a pulse train where the pulse duration is proportional to the distance between the transducer and the nearest sound-reflecting object.

An ultrasonic transducer is pictured in [Figure 27-2](#). Its internal components are shown in [Figure 27-3](#).

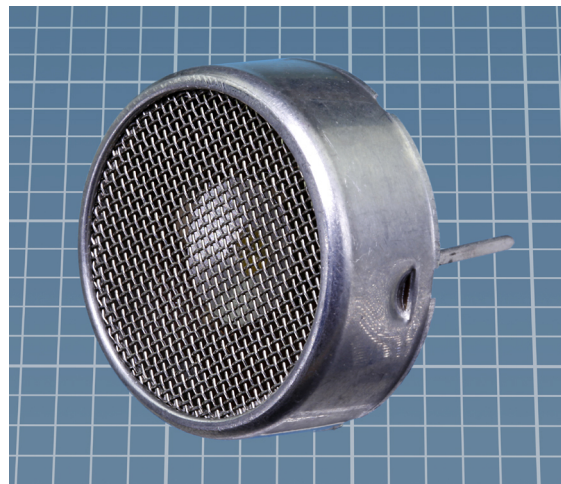


Figure 27-2. The exterior of an ultrasonic transducer.

Submersible ultrasonic transducers may be used in cleaning systems, where they agitate a liquid that dislodges dirt or debris. Ultrasonic transducers are also used in echo-sounding and sonar equipment with marine applications.

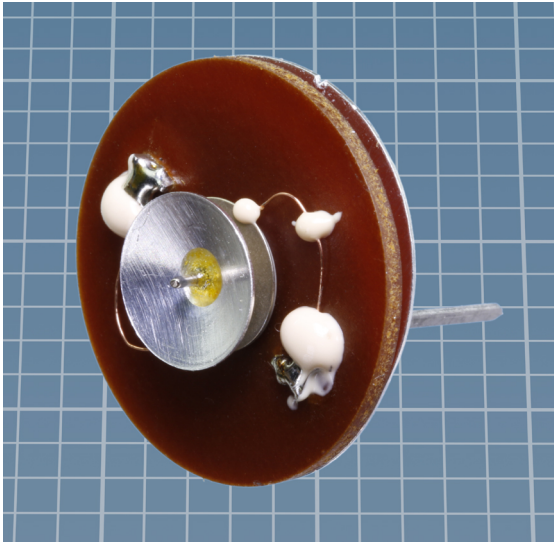


Figure 27-3. Inside an ultrasonic transducer, a small aluminum cone is the sound radiating element. The white blobs are adhesive to secure the thin wires.

Formats

Some transducers are available in surface-mount format, measuring about 0.5" square or less. Because the resonant frequency is related to the size of the component, surface-mount transducers usually generate a high-pitched beep.

Values

Frequency Range

Audio frequency is measured in Hertz, abbreviated Hz, named after Heinrich Rudolf Hertz, the first scientist to prove the existence of electromagnetic waves. The H in Hz is capitalized because it refers to a real name. One thousand Hertz can be written as 1 kiloHertz, almost always abbreviated as 1kHz (note that the k is lowercase).

The human ear is often described as being able to detect sounds between 20Hz and 20kHz, although the ability to hear sounds above 15kHz is relatively unusual and diminishes naturally with age. Sensitivity to all frequencies can be impaired by long-term exposure to loud noise.

The most common frequencies applied to audio transducers range between 3kHz and 3.5kHz.

Piezoelectric elements are inefficient for generating sounds below 1kHz, but electromagnetic transducers are better able to generate lower frequencies. Their response curve can be approximately flat to frequencies as low as 100Hz.

Sound Pressure

Sound pressure can be measured in Newtons per square meter, often abbreviated as Pa. Newtons are units of force, while Pa is an abbreviation of Pascals.

The *sound pressure level* (SPL) of a sound is not the same as its sound pressure. SPL is a logarithmic value, to base 10, in units of decibels (dB), derived from the pressure of a sound wave relative to an arbitrary reference value, which is 20 micro-Pascals (20μPa). This is the agreed minimum threshold of human hearing, comparable to a mosquito at a distance of three meters. It is assigned the level of 0dB.

Because the decibel scale is logarithmic, a linear increase in the decibel level of a sound does not correspond with a linear increase in actual sound pressure:

- For each additional 6dB in the SPL, the actual sound pressure approximately doubles.
- For each additional 20dB in the SPL, the actual sound pressure is multiplied by 10.

Bearing in mind that 0dB corresponds with the reference sound pressure of 20μPa, an SPL of 20dB represents a sound pressure of 200μPa (that is 0.0002Pa), and so on.

Many tables show an estimated decibel level for various noise sources. Unfortunately, these tables may contradict each other, or may fail to mention the distance at which a sound is measured. [Figure 27-4](#) shows estimates derived by averaging eight similar tables. It should be viewed as an approximate guide.

Decibels	Noise Example
140	Jet engine at 50 meters
130	Threshold of pain
120	Loud rock concert
110	Automobile horn at 1 meter
100	Jackhammer at 1 meter
90	Propeller plane 300 meters above
80	Freight train at 15 meters
70	Vacuum cleaner
60	Business office
50	Conversation
40	Library
30	Quiet bedroom
20	Leaves rustling
10	Calm breathing at 1 meter
0	Auditory threshold

Figure 27-4. Approximate decibel values for some sound sources (averaged from a selection of eight similar charts).

Sometimes the claim is made that an increase of +10 on the decibel scale will correspond with a subjective experience that the noise is “twice as loud.” Unfortunately, this statement cannot be quantified.

Weighted Sound Values

Subjective assessment of sound is complicated by the nonlinear frequency response of the human ear, which causes some frequencies to seem “louder” than others, even though their sound pressure is the same. The frequency weighting of the ear can be determined by playing a reference tone of 1kHz at 20dB and then doing an A-B comparison with a secondary tone at another frequency, asking the subject to adjust the gain of the secondary tone up or down until the two tones seem equally loud.

This procedure is performed for a range of frequencies. The test is then repeated with a louder 1kHz reference tone, at 30dB. Repetitions continue to a final reference tone of 90dB.

The resulting curves are known as *equal-loudness contours*. An averaged set, from multiple sources, has become an international standard with ISO number 226:2003. The curves shown in Figure 27-5 are derived from that standard. The curves show that the sound pressure of lower frequencies must be boosted by a significant amount to sound as loud as a 1kHz frequency, while a frequency around 3kHz must be reduced slightly, because it tends to sound louder than all others.

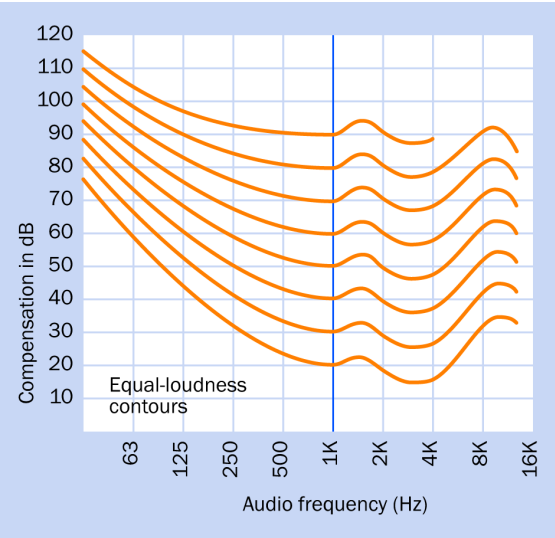


Figure 27-5. Equal loudness contours derived from ISO 226:2003. See text for details.

Although the accuracy of equal-loudness contours is controversial, they have been the basis of a widely used weighting system to adjust dB values to represent subjective perceptions of loudness. This *A-weighting* system remains the best-known and most widely applied audio standard in the United States, even though it has been criticized for assigning too little value to sounds that are brief in duration. If a sound level is expressed in dBA, it is A-weighted, meaning that the sounds to which the ear is least sensitive are assigned a value that is lower than their measured value. Thus, a tone of 100Hz has a dBA value about 20dB lower than its dB value, because the human ear is relatively insensitive to

low-pitched sounds. dBA values are used in regulations that limit noise in the work place and other environments.

Unweighted Values

If sound intensity is expressed in dB SPL, it is a measurement of the actual Sound Pressure Level and has not been adjusted with the A-weighting system. A graph of unadjusted dB SPL values will display low frequencies as if they are more intense than the ear will perceive. In practical terms, subjective perception of low-end rolloff will be even more severe than the graph makes it appear.

If sound intensity is expressed merely in dB, probably it is unweighted and should be considered as dB SPL.

From a practical point of view, when choosing a tone for a transducer, a 500Hz tone may sound relatively mellow and not subjectively loud. A 3.5kHz tone can be a good attention-getting signal, as the ear is most sensitive in that range.

Transducers generally have a sound pressure rating in dB SPL ranging between 65dB SPL to 95dB SPL, with just a few products that can make more or less noise.

Measurement Location

The sound pressure from an audio alert will naturally diminish if the measurement point moves farther away. Therefore, any rating in decibels should be expressed with reference to the distance at which the measurement is made.

Measurement locations may be expressed in centimeters or inches, and may vary from 10cm to 1 meter, even in datasheets for different devices from the same manufacturer. If the measurement distance doubles, the SPL diminishes by approximately 6dB.

Limitations

A piezoelectric transducer is not intended as a sound reproducer, and does not have a smooth or flat frequency response. The curve for the Mal-

lory PT-2040PQ is not unusual, reproduced in [Figure 27-6](#). This component measures about 3/4" in diameter, is rated for 5VDC, and uses only 1.5mA to generate 90dB (measured at a distance of 10cm). Like many piezoelectric audio devices, its response peaks around 3500kHz and diminishes above and below that value, especially toward the low end. While it is perfectly adequate as a "beeper," it will not reproduce music successfully.

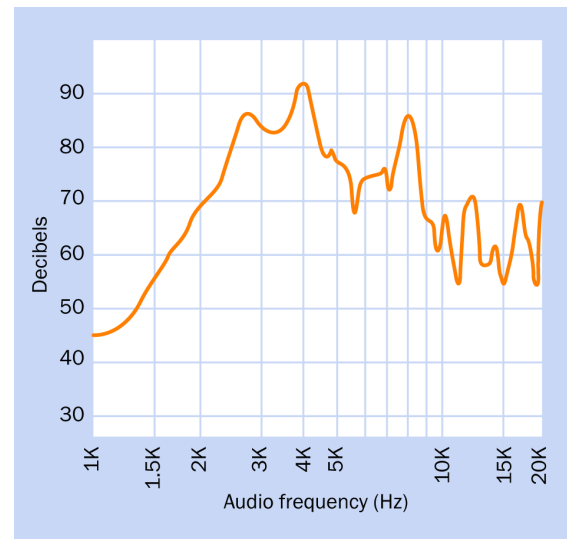


Figure 27-6. The frequency response from a typical small piezoelectric transducer.

An electromagnetic transducer is better able to generate low frequencies than a piezoelectric transducer. It has a low impedance that may be suitable in some circuits. However, it is slightly heavier than a comparable piezoelectric transducer, uses much more power, and as an AC device containing a coil, it can create electromagnetic interference or may cause fluctuations in the circuit as an inductive load. It is also vulnerable to magnetic interference from elsewhere, while a piezoelectric transducer is not.

While an electromagnetic transducer can be used to reproduce speech or music, and will do a better job than a piezoelectric transducer, its

performance will still be dissatisfying. A miniature speaker is more appropriate for the task.

Voltage

Transducers are typically designed to work with voltages ranging from 5VAC to 24VAC. The ceramic wafer in a piezoelectric transducer usually cannot withstand voltages much above 40VAC, and its sound output will not increase significantly above 30VAC.

Current

Typical piezoelectric transducers use less than 10mA and generate negligible heat. An electromagnetic transducer may draw as much as 60mA.

How to Use It

Appropriate Sound Intensity

An alert should be chosen with reference to the environment in which it will be used. To be easily heard, it should be at least 10 dB louder than ambient background noise.

Volume Control

Sound pressure can be lowered by reducing the voltage. Because a transducer does not consume much current, a *trimmer* can serve as a volume control. Alternatively, a rotary switch with a set of fixed-value resistors can select preset sound values.

AC Supply

Although a transducer is an AC device, it is unlikely to be designed for voltage that fluctuates positively and negatively either side of a neutral value. Typically it is intended for voltage that fluctuates between 0V (ground) and the rated positive value of the power supply, and its pins, wires, or terminals are usually marked accordingly. If it has wire leads, the red lead should be connected to the more-positive side of the supply. If it has pins, the longer pin should be more positive.

The alternating signal for a transducer can be supplied by any simple oscillator or astable multivibrator circuit. For a given peak voltage, a square wave will generate a louder signal than a sinusoidal wave. A simple 555 timer circuit can be used, with a second monostable timer to limit the duration of the beep if necessary. An astable 555 can be used to test the transducer and select the audio frequency that sounds best.

Self-Drive Transducer Circuit

If a transducer has three wires or pins, it is probably a *self-drive* type. The datasheet may identify its inputs as M, G, and F, meaning Main, Ground, and Feedback. The Feedback terminal is connected with a section of the diaphragm which vibrates 180 degrees out of phase with the Main terminal. This facilitates a very simple external drive circuit, such as that in [Figure 27-7](#), where the frequency is determined by the transducer's resonant frequency.

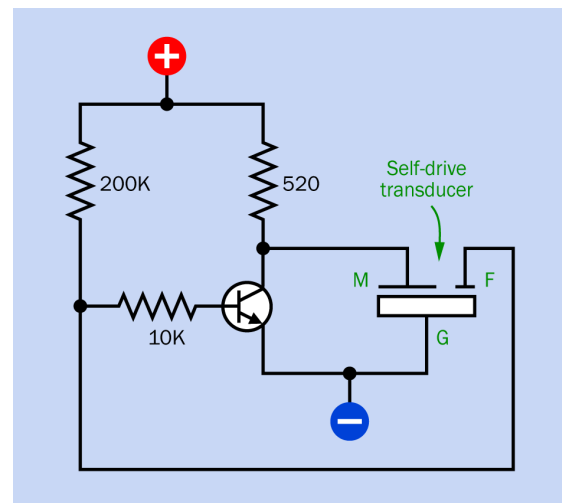


Figure 27-7. A circuit to control a self-drive type of piezoelectric transducer.

What Can Go Wrong

Overvoltage

Mallory Sonalerts, one of the largest producer of piezoelectric alerts, states that in the “vast ma-

jority” of returned products, the failure was caused by excessive voltage, often in the form of a transient voltage spike.

Leakage

If an alert makes a small, low-volume sound when it is supposed to be off, some current is leaking through it. Less than 1mA may be enough to cause this problem. According to one manufacturer, it can be fixed by placing a 30V Zener transient voltage suppressor diode in series with the alert, or by wiring a small **incandescent lamp** in parallel with the alert.

Note that when the alert is activated, the full supply voltage will be seen at the lamp.

Component Mounting Problems

Some alerts are packaged with mounting holes, but many are not. Those with pins can be soldered into a board, but those without must be glued in place or inserted into a cavity from which they cannot shake loose. Silicone adhesive is recommended, but care must be taken to avoid any of it dripping into the alert before it sets.

Moisture

If an alert will be used in a location where it is vulnerable to moisture, it should be of a type that is sealed against the environment. Even a sealed unit should ideally be oriented so that it faces slightly downward.

Transducer-Indicator Confusion

Externally, a transducer and an indicator often look identical, and some of them are not marked with a manufacturer’s part number. Damage can be caused by applying DC to a transducer or AC to an indicator. If both types of parts are kept in inventory, they should be carefully labeled.

Connection with a Microcontroller

A piezoelectric transducer can be driven by a microcontroller, but an electromagnetic transducer is not appropriate in that role, because of its relatively higher current consumption and its behavior as an inductive load.

audio indicator

An **audio indicator** is defined here as a noise-creation device that generates a simple tone or series of tones. Unlike a **transducer**, which requires an external source of AC to determine its audio frequency, an indicator contains its own electronics and requires only a DC power supply. Either of these components is often described as a *beeper* or *buzzer*.

While piezoelectric alerts formerly used crystals, only the more modern piezoelectric type that uses a ceramic wafer will be considered here.

OTHER RELATED COMPONENTS

- **transducer** (see [Chapter 27](#))
- **headphone** (see [Chapter 29](#))
- **speaker** (see [Chapter 30](#))

What It Does

When DC power is applied to an **audio indicator**, in its simplest form it creates a continuous or intermittent tone of a fixed frequency. This is called an *alert*.

Audio alerts are used in microwave ovens, washer/dryers, automobiles, gasoline pumps, security devices, toys, phones, and many other consumer products. They are often applied with touch pads, to provide audio confirmation that a tactile switch has been pressed.

A few indicators are programmed to create a two-tone sound, or multiple-tone sequences.

See [Figure 27-1](#) in the previous entry for an assortment of schematic symbols that may be used to represent either an alert or a transducer.

How It Works

A circular diaphragm is glued at its edges inside a cylindrical plastic enclosure, usually measuring

from around 0.5" to 1.5" in diameter. The enclosure is sealed at the bottom but has a small hole at the top, so that sound can emerge from the upper side of the diaphragm without being partially cancelled by sound of opposite phase that is emitted from the underside of the diaphragm. The enclosure also contains electronics to generate one or more audio tones, and amplifies the sound by resonating with it, in the same way that the body of a guitar or violin amplifies a note being played on the strings.

A PUI XL453 piezoelectric audio indicator is pictured in [Figure 28-1](#), fully assembled on the right, and with its circuit board and diaphragm removed on the left. This indicator creates a pulsed tone at 3.5kHz with a sound pressure of 96dB. It draws 6mA at 12VDC and measures approximately 1" in diameter.

For more information about the measurement of sound frequency and pressure, see "[Frequency Range](#)" on [page 251](#) and "[Sound Pressure](#)" on [page 251](#) in the previous entry.

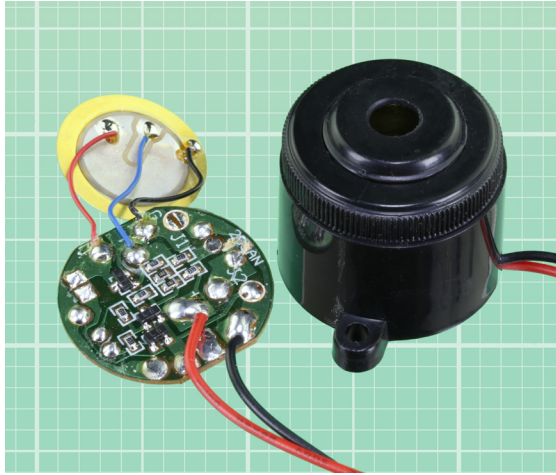


Figure 28-1. A typical piezoelectric audio indicator.

Externally, an audio indicator may be indistinguishable from a transducer. However, internally, an indicator is almost always a piezoelectric device, in which a *ceramic wafer* is mounted on a thin brass diaphragm. The term *piezo* is derived from the Greek *piezein*, which means to squeeze or press.

A transducer (described in the previous entry) is a piezoelectric or electromagnetic alert that does not usually contain its own circuitry and must be driven by an external source of AC, which establishes the audio frequency.

The distinction between an indicator and a transducer is often unclear in parts catalogs, where all alerts may be identified as *buzzers*, even though they mostly beep rather than buzz.

Audio Frequency

For a discussion of audio frequency, see “[Frequency Range](#)” on page 251 in the previous entry.

History

Probably the earliest form of electrically activated alert was the door bell, in which a 6VDC battery-powered solenoid pulled a spring-loaded lever terminating in a small hammer. The hammer struck the bell, but the motion of the

lever also opened a pair of contacts, cutting off power to the solenoid. The lever sprang back to its rest position, which closed the contacts and repeated the cycle so long as power was supplied through an external pushbutton.

Subsequent systems used a small loudspeaker powered by AC house current through a step-down transformer. This created a buzzing sound and may have been the origin of the term “buzzer.”

Small components that made a beeping sound only became common when digital equipment required a simple, cheap way to confirm user input or attract attention to the status of a device.

Variants

Sound Patterns

Because an audio indicator contains its own electronics, the manufacturer has the freedom to create various patterns of sound output.

The default is a steady tone. Other common variants include an intermittent tone and a dual tone that fluctuates rapidly between two frequencies. This is sometimes referred to as a *si-ren*. A few variants can generate an output pattern consisting of several tones in sequence, or effects such as warbling or whooping sounds, which are used mainly in alarm systems.

Formats

Some audio indicators are available in surface-mount format, measuring 1/2” square or less. Because the resonant frequency is related to the size of the component, surface-mount alerts usually make a high-pitched beep.

Panel-mount and board-mount formats range from about 1/2” to 1.5” in diameter. A small audio alert designed to be mounted on a circuit board is shown in [Figure 28-2](#), with its top removed on the right to expose the brass diaphragm glued around the edges. The same component is shown with its plastic enclosure removed completely in [Figure 28-3](#).

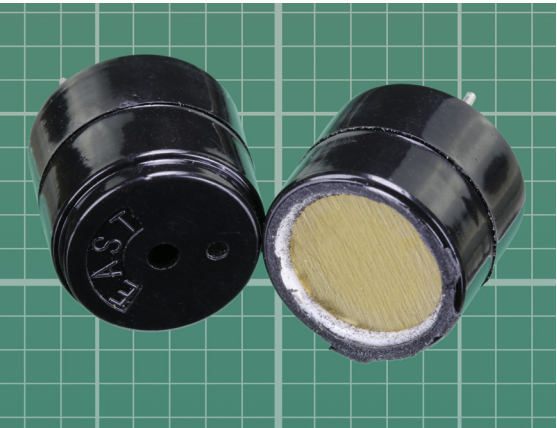


Figure 28-2. An audio indicator approximately 0.5" in diameter, partially disassembled on the right, revealing its brass diaphragm.

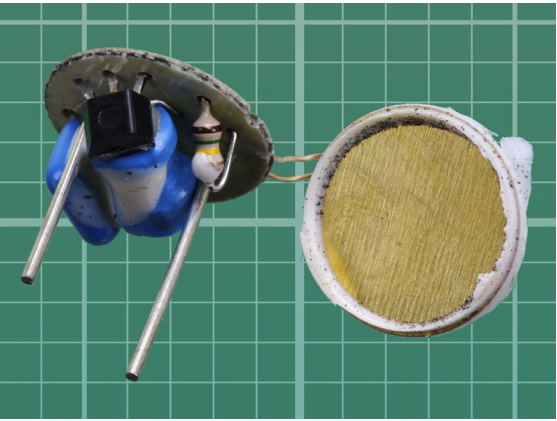


Figure 28-3. The same indicator from the previous photograph, with its enclosure completely removed.

Values

For an explanation and discussion of sound pressure and its measurement in decibels, see “[Sound Pressure](#)” on page 251 in the previous entry.

Audio indicators generally have a sound pressure rating in dB SPL ranging between 65 dB SPL to 95 dB SPL, with just a few products that make more or less noise. At 120 dB and above, most products are packaged as alarm sirens ready for installation, often with a small horn attached. Their power consumption can be 200 mA or

more, and they are many times the price of a simple indicator designed for circuit-board mounting.

Voltage

An audio indicator containing its own electronics will almost always be rated somewhere in the range from 5 VDC to 24 VDC. Sirens intended for use with burglar alarms are often designed for 12 VDC or 24 VDC, as these are popular values for security systems with battery backup. However, in addition to a rated voltage, a datasheet may specify a wide range of acceptable operating voltages. For example, an indicator with a *rated* voltage of 12 VDC may have an *operating* voltage of 3 VDC to 24 VDC. Naturally, the sound intensity will vary with the voltage, but not as much as one might assume. The graph at [Figure 28-4](#) shows that the sound output from an alarm, measured in decibels, increases by only 8 dB when voltage increases by almost a factor of five. Of course, the decibel scale is not linear, but human perception of sound is not linear, either.

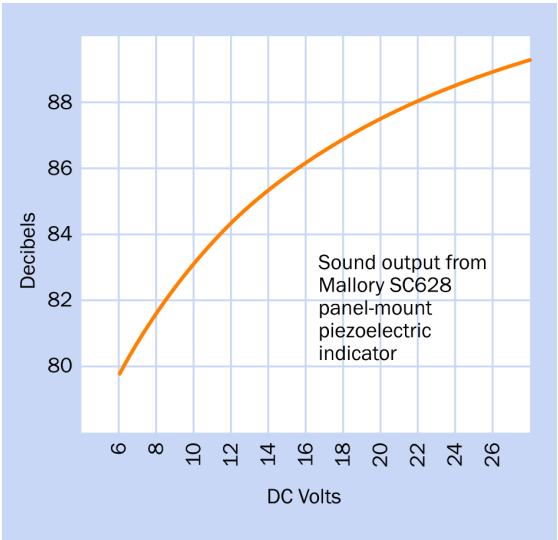


Figure 28-4. Variation of sound output relative to voltage, in a commonly used piezoelectric indicator.

Current

Typical piezoelectric indicators use less than 10mA (often as little as 5mA) and generate negligible heat.

Frequency

The most common frequencies for indicators range between 3kHz and 3.5kHz. Piezoelectric elements are inefficient for generating sounds below 1kHz.

Duty Cycle

Piezoelectric alerts generate very little heat and can be run on a 100% duty cycle.

If an alert will be pulsed briefly, the minimum pulse time is 50ms. A shorter duration will merely generate a clicking sound.

How to Use It

Appropriate Sound Intensity

An indicator should be chosen with reference to the environment in which it will be used. To be easily heard, it should be at least 10 dB louder than the ambient background noise.

Volume Control

Sound intensity can be reduced by reducing the voltage. Because an indicator does not consume much current, a trimmer can serve as a volume control. Alternatively, a rotary switch with a set of fixed-value resistors can select preset sound values.

However, in many indicators, variations in voltage may have relatively little effect on sound output, as shown in [Figure 28-4](#).

Wiring

An indicator requires DC voltage. Because the indicator contains a transistor, polarity of the power supply is important. If the indicator has leads attached, the one intended for connection to the positive side of the power supply will be red. If it has pins, the longer pin will be for the positive connection.

What Can Go Wrong

The potential problems in an indicator are the same as those for a transducer. See [“What Can Go Wrong” on page 254](#) in the previous entry.

headphone

The term **headphone** is used here to include almost any device that fits into or over the ear for the purpose of sound reproduction. (*Hearing aids* are not included.) Because headphones are used in pairs, the term is usually pluralized.

The term *phones* is fairly common as a colloquial diminution of *headphones* but is not used here.

An *earphone* used to be a single sound reproduction device designed for insertion into the ear, but has become rare. Pairs of *earbuds* are now common.

Because this encyclopedia assigns more emphasis to electronic components than to consumer products, this entry provides only a superficial overview of fully assembled headphones, and deals more with the drivers inside them, their principles of operation, and the general topic of sound reproduction.

OTHER RELATED COMPONENTS

- **transducer** (see [Chapter 27](#))
- **speaker** (see [Chapter 30](#))

What It Does

A headphone converts fluctuations of an electric signal into pressure waves that the human ear perceives as sound. It can be used for reproduction of music for entertainment purposes, or for speech in telecommunications, broadcasting, and audio recording.

Two symbols for headphones are shown in [Figure 29-1](#). The symbol on the left shows a single headphone or earphone; when this symbol is flipped horizontally, it can represent a microphone. The pictographic symbol on the right has been used for many decades, but is still often found in schematics.



Figure 29-1. Schematic symbols for a single earphone or headphone (left) and a pair of headphones (right).

How It Works

Audio Basics

Sound is transmitted as pressure waves through a medium that is usually air but can be a gas, fluid, or solid. The speed of transmission will vary with the density and other attributes of the medium. Small hairs in the inner ears, known as *cilia*, vibrate in sympathy with pressure waves and transmit nerve impulses to the brain, which interprets the impulses as sound.

Three quantities describe the propagation of any type of wave, including a sound wave: its frequency (customarily represented with letter f), its speed of propagation (represented with letter v , for velocity), and its wavelength from peak-to-peak (represented by the Greek letter lambda, which appears as this λ symbol).

The relationship is defined by a very simple equation:

$$v = \lambda * f$$

Velocity is usually measured in meters per second, wavelength in meters, and frequency in Hertz, abbreviated Hz. One cycle per second is 1Hz. The H is always capitalized, as it refers to the name of Heinrich Rudolf Hertz, the first scientist to prove the existence of electromagnetic waves. One thousand Hertz can be written as 1 kilo-Hertz, almost always abbreviated as 1kHz (note that the k is lowercase).

The human ear is often described as being able to detect sounds between 20Hz and 20kHz, although the ability to hear sounds above 15kHz is relatively unusual and diminishes naturally with age. Sensitivity to all frequencies can be impaired by long-term exposure to loud noise.

Naturally occurring sounds can be converted to fluctuations in voltage by a **microphone**, which will be found listed as a [sensor](#) in Volume 3 of this encyclopedia. Artificial sounds can be generated as voltage fluctuations by oscillators and other electronic circuits. In either case, the output fluctuations can range between an upper limit set by a positive supply voltage and a lower limit established by electrical ground (which is assumed to be 0 volts). Alternatively, the fluctuations can range between the positive supply voltage and an equal and opposite negative supply voltage, with 0V lying midway between the two. This option can be less convenient electrically but is a more direct representation of sound, because sound waves fluctuate above and below ambient air pressure, which can be considered analogous to a ground state.

The concept of positive and negative sound waves is illustrated in [Figure 29-2](#) (originally published in the book [Make: More Electronics](#)).

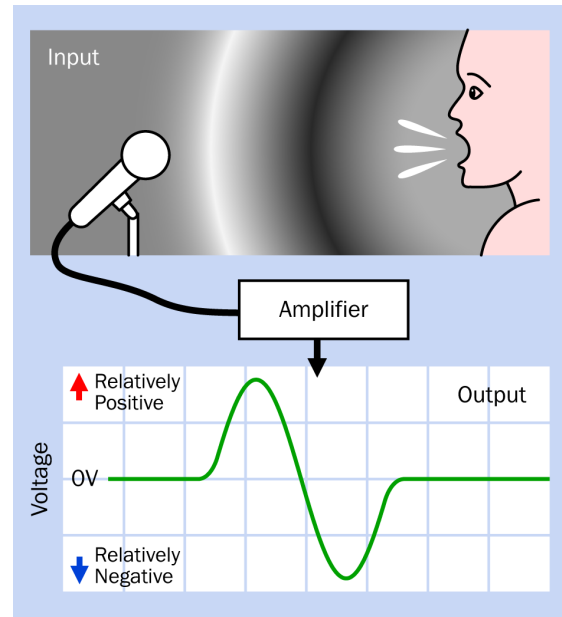


Figure 29-2. The use of positive and negative voltages to represent a wave of high sound pressure followed by a trough of lower pressure.

The topic of sound amplification is explored in detail in the entry on [op-amps](#) in [Chapter 7](#).

A headphone inverts the function of a microphone by converting electricity back into air-pressure waves. This is done electromagnetically (moving a diaphragm in response to an electromagnet) or electrostatically (moving a membrane in response to electrostatic force between two charged electrodes).

Variants

Moving Coil

The most enduringly popular type of headphone uses a coil attached to a [diaphragm](#). This is known as a [moving-coil](#) headphone, as the coil moves with the diaphragm. It can also be referred to as having a [dynamic driver](#) or [dynamic transducer](#), “dynamic” referring to the movement of the coil.

The moving-coil concept is illustrated in [Figure 29-3](#). The coil slides into a deep, narrow, circular slot in a magnet that is attached to the plastic frame of the headphone. The diaphragm is supported at its edges by a flexible rim. Variations in current passing through the coil create a fluctuating magnetic field that interacts with the field of the fixed magnet, causing the diaphragm to move in and out. A very similar configuration is used in many **loudspeakers**. Detail modifications may be made to increase efficiency, reduce production costs, or enhance sound quality, but the principle remains the same.

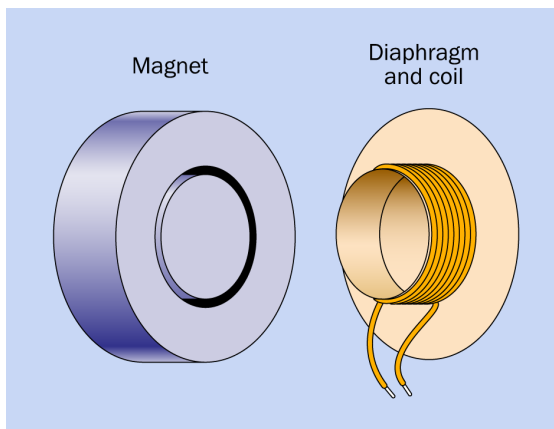


Figure 29-3. The basic elements of a moving-coil headphone.

The internal element of a headphone is shown in [Figure 29-4](#). A plastic diaphragm is visible, measuring slightly less than 2" in diameter. The magnet and coil are concealed underneath.

The element in the previous figure is normally enclosed in an assembly such as the one in [Figure 29-5](#), which incorporates a soft padded rim to rest upon the ear.

In an effort to achieve a more balanced frequency response, some designs use two moving-coil drivers in each headphone, optimized for low frequencies and high frequencies, respectively.

Earbuds, described after the next section, often use a miniaturized version of the moving-coil design.



Figure 29-4. The sound reproducing element removed from a headphone.

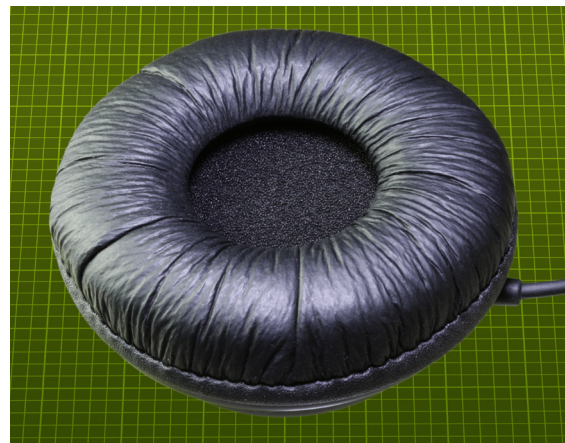


Figure 29-5. The sound reproducing element from the previous figure is normally packaged inside this enclosure.

Other Types

Electrostatic headphones use a thin, flat diaphragm suspended between two grids that function as electrodes. A fluctuating potential between the grids, coupled with a reverse-phase voltage on the diaphragm, will vibrate the diaphragm, generating pressure waves. A relatively high voltage is necessary to achieve this, anywhere from 100V to 1,000V, supplied through a conversion unit between the headphones and an amplifier. Electrostatic headphones are

known for low distortion and an excellent high frequency response, at some extra cost.

Electret headphones work on a similar principle, except that the membrane is permanently charged, and a high voltage is not required. Electret headphones tend to be small, inexpensive, and not of high sound quality.

A *balanced armature* design, often referred to by the acronym *BA*, uses a pivoting magnet that is claimed to increase efficiency while reducing stress on a diaphragm. BA drivers can be extremely compact, contained within a sealed metal enclosure measuring less than 10mm x 10mm x 5mm. They are commonly used in conjunction with *in-ear* earphones, described in the next section.

Mechanical Design

Circumaural headphones use large soft pads to encircle the ear and block external noise. Their size tends to make them heavy, requiring a well-designed headband to provide comfortable support. *Supra-aural* headphones are smaller and lighter, resting on the ears instead of enclosing them. They cannot exclude ambient noise, and may have inferior bass response compared with the circumaural type.

Open-back headphones, also known as *acoustically transparent*, are favored by some audiophiles because their vented outer surfaces are thought to create a more natural sound, similar to that of a **speaker**. The open backs naturally allow ambient noise to intrude, but also allow the sound generated by the headphones to be heard by others in a room. *Closed-back* headphones contain their sound and provide more insulation against ambient noise.

Earbuds rest just within the outer folds of the ear, facing inward like a pair of tiny **speakers**. They are easily dislodged and provide very little insulation against ambient noise. Their use became common after the introduction of Apple's iPod. A pair of earbuds, one of them with its plastic cover removed, is shown in [Figure 29-6](#).

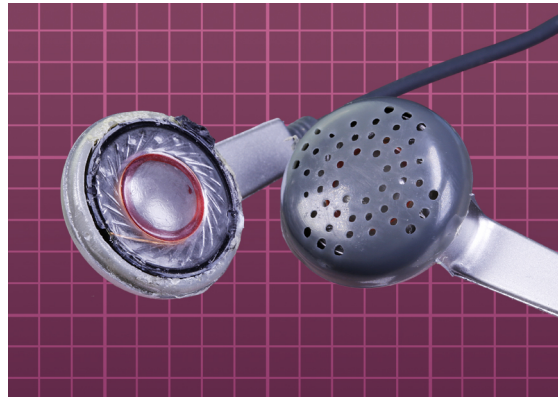


Figure 29-6. A pair of earbuds, one with its cover removed to show the sound reproducing element, which closely resembles the diaphragm in a full-sized headphone.

In-ear headphones are designed for insertion into the ear canal, often using a soft sheath that conforms with the ear like an earplug. This sheath is disposable for hygienic reasons, and because it may lose some of its plasticity with use. It excludes most ambient noise, and by minimizing the air gap between the driver of the headphone and the ear drum enables a high quality of sound reproduction.

In-ear headphones are also known as *in-ear monitors*, *IEMs*, *ear canal headphones*, *earphones*, and *canalphones*. A pair of in-ear headphones is shown in [Figure 29-7](#), one of them with its foam sheath removed. The rectangular silver-colored object in the headphone on the left contains a transducer to create sound pressure.

A *headset* consists of one or two headphones plus a flexible microphone that extends to the promiximity of the mouth of the user.

Noise-cancelling headphones, popularized by Bose, monitor external noise with a built-in microphone and generate sound of opposite phase, to provide some cancellation. They are particularly effective on jet aircraft, where background noise tends to be consistent.



Figure 29-7. A pair of in-ear headphones, supplied with disposable foam plugs that conform flexibly with the ear canal. The headphone on the left is shown with its plug removed.

Although an *earphone* is almost obsolete, it is still obtainable from specialty suppliers. It has a high impedance, making it suitable for use with a *crystal-set radio*. An earphone is pictured in Figure 29-8.



Figure 29-8. A vintage earphone of the type suitable for use with a crystal-set radio.

Values

Intensity

Sound pressure is measured in decibels. For a complete explanation and discussion of weighted and unweighted decibel scales, see “[Sound Pressure](#)” on page 251 in the **transducer** entry.

Frequency Response

A plot of sound pressure as a function of frequency shows the *frequency response* of a headphone. Measuring the sound pressure meaningfully is a challenge, because the ear canal will add coloration to the sound and can amplify some frequencies while masking others. Ideally, measurement should be done at the ear drum, but this is not feasible. Consequently, high-end headphones are evaluated by making sound measurements inside simulated ear canals in a dummy human head.

A comparison between a high-quality \$500 audio product and a transducer that is sold as a component for less than \$1 illustrates the difference in frequency response; see Figure 29-9. The Sennheiser headphones have a smooth response that rises toward the low end, compensating for the lack of bass response that tends to be a problem in headphones, and the relative insensitivity of the human ear to low frequencies. The fluctuations at the high end are within about 5dB.

By comparison, the Kobitone emphasizes the range between 3kHz and 4kHz because its primary task is to be heard, and these are the frequencies where human hearing is most sensitive. Its low-frequency response trails off (although is still much better than that of a piezoelectric transducer, where the low response typically diminishes by 40dB to 50dB). The low-frequency output of the Kobitone is actually impressive bearing in mind that the component is only 9mm in diameter. It draws 60mA at 5VAC.

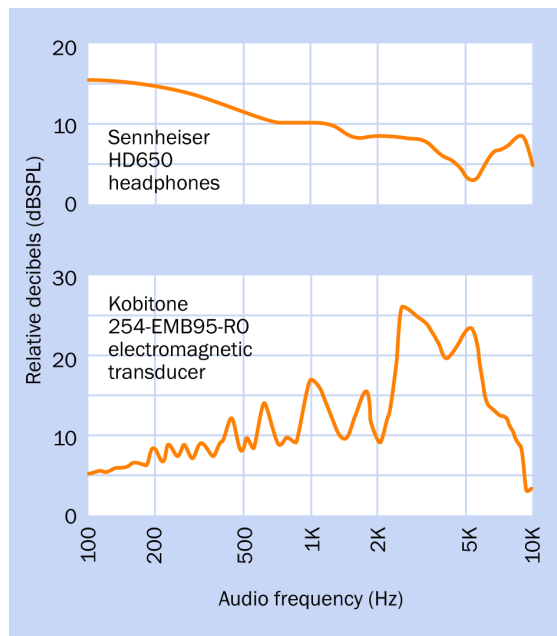


Figure 29-9. Comparison between frequency responses of a \$1 electromagnetic transducer intended as an audio alert and a \$500 pair of headphones intended for sound reproduction. The upper graph is derived from a review online at headroom.com; the lower graph is from the manufacturer's datasheet.

Some manufacturers of audio reproduction equipment prefer not to supply frequency response curves. Instead they may claim, for example, that the frequency response of a product ranges from 100Hz to 20kHz. This claim means very little unless it is accompanied by a range of sound pressure levels. If the frequency response is consistent within a range of, say, plus-or-minus 5dB, this may be acceptable. If the range is plus-or-minus 20dB, it is not acceptable. The ability to reproduce a high note or a low note is not useful if the sound is too faint to hear.

Distortion

The *total harmonic distortion* (THD) of any audio equipment measures its tendency to add spurious *harmonics* of a single frequency. If headphones are required to reproduce a pure 1kHz sinewave, they will also tend to create an additional 3kHz tone that is an artifact. This can be

caused by the mechanical behavior of a vibrating diaphragm. The human ear recognizes distortion as a fuzzy or rasping sound. A square wave theoretically contains all the harmonics that are odd multiples of its fundamental frequency, and sounds extremely distorted.

THD should be less than 1% in good-quality audio devices.

Impedance

The electrical impedance of headphones is relevant in that it should match the output specification of the amplifier that drives them.

What Can Go Wrong

Overdriving

Headphones can be damaged by overdriving them. Because a low frequency requires larger excursions of a diaphragm to transmit the same energy as a high frequency, headphones are especially vulnerable to being damaged by bass at high volume.

Hearing Damage

Human hearing can be damaged by prolonged listening to headphones at a high volume. Some controversy remains regarding an acceptable limit for sound pressure.

Mismatched Impedance

If the impedance of headphones does not match the output of the amplifier driving them, distortion or a skewed frequency response can result. This is known as *mismatching*.

Incorrect Wiring

In most consumer products, a pair of headphones will share a common ground. While the connections in a typical three-layer jack plug have been standardized, hand-wired repairs or extensions should be tested carefully. Incorrect wiring will cause unpredictable results.

speaker

The term **speaker** is a diminution of *loudspeaker*. The full word is now so rarely used, some catalogs do not recognize it as a search term. This encyclopedia acknowledges contemporary usage by using **speaker** rather than *loudspeaker*.

A fully assembled consumer product can be referred to as a “speaker,” but it also contains one or more individual components that are described as “speakers.” To resolve this ambiguity, referring to the components as *drivers* is helpful, but this practice can create more ambiguity because other types of components are also called “drivers.” The only real guide to the meaning of *speaker* is the context in which it is used.

For the purposes of this entry, a speaker is a *sound reproduction device*, distinguished from a typical electromagnetic **transducer** by being larger and more powerful, with a more linear frequency response. A transducer may be used as a *noise-creating device* to provide an alert, informing the user of the status of a piece of equipment. Because some speakers have been miniaturized for use in handheld products, they may be used as transducers, allowing some overlap between the two categories.

Because this encyclopedia assigns more emphasis to electronic components than to consumer products, this entry provides only a superficial overview of fully assembled speakers, and deals more with the drivers inside them, their principles of operation, and the general topic of sound reproduction.

OTHER RELATED COMPONENTS

- **headphone** (see [Chapter 29](#))
- **transducer** (see [Chapter 27](#))

What It Does

A speaker converts fluctuations of an electric signal into pressure waves that the human ear perceives as sound. It can be used for entertainment purposes or to provide information in the form of spoken words or distinctive sounds (as in the case of a miniature speaker in a cellular phone, playing a ring tone).

The internationally accepted schematic symbol for a speaker is shown in [Figure 30-1](#).



Figure 30-1. Only one symbol exists to represent a speaker. This is it.

How It Works

For a summary of basic concepts and terminology relating to sound and its reproduction, see “[Audio Basics](#)” on [page 261](#) in the previous entry.

Construction

A speaker contains a *diaphragm* or *cone* with a coil attached to it. Fluctuations of current through the coil interact with a permanent magnet, causing the speaker to emit pressure waves that are proportionate with the current. The design is similar in concept to that of a headphone, shown diagrammatically in [Figure 29-3](#). The primary difference is that a speaker of around 2" or more will use a cone rather than a flat diaphragm. The cone shape is more rigid and creates a more directional sound.

A 2" speaker rated for 1/4W with a 63Ω coil is shown in [Figure 30-2](#), undamaged on the left but with its cone cut out on the right. The neck of the cone, which is normally inserted in the circular groove in the speaker magnet, is shown with the inductive coil wrapped around it.

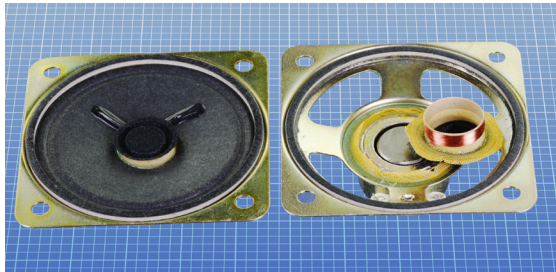


Figure 30-2. On the left is a 2" speaker. On the right, its cone has been cut away to reveal the magnet, with a circular groove in it. The neck of the cone, which normally slides into the groove, is shown removed.

A speaker with a cone 4" in diameter is shown from the rear in [Figure 30-3](#).

A miniature surface-mount speaker is shown from the front and from the rear in [Figures 30-4](#) and [30-5](#). It measures just under 0.4" diameter and was made for Motorola. Its power rating is 50mW.

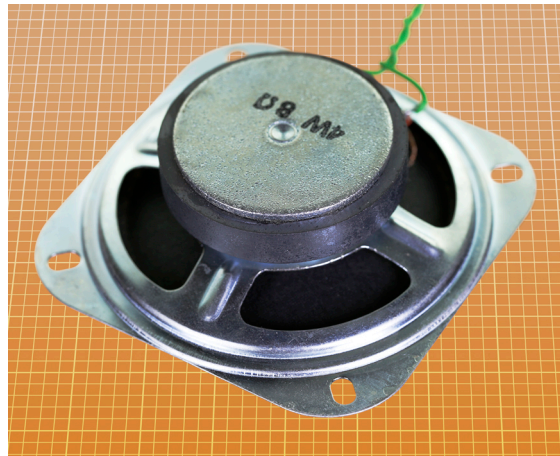


Figure 30-3. The back side of a speaker with a cone measuring approximately 4" in diameter. Its magnet is the large round section that is uppermost. This unit is rated for 4W and has an impedance of 8Ω.

A speaker designed for a cellular phone is shown in [Figure 30-6](#). Note the close resemblance in design to the driver used in an earbud, shown in [Figure 29-6](#) in the previous entry.

In the past, speaker cones were made from tough, fibrous paper. Modern cones are more likely to be plastic, especially in small sizes.



Figure 30-4. Front view of a surface-mount speaker measuring less than 0.4" diameter.

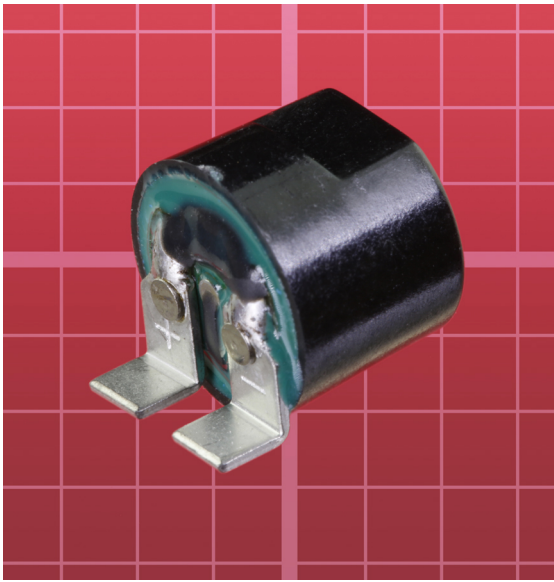


Figure 30-5. Rear view of the speaker shown in the previous figure.

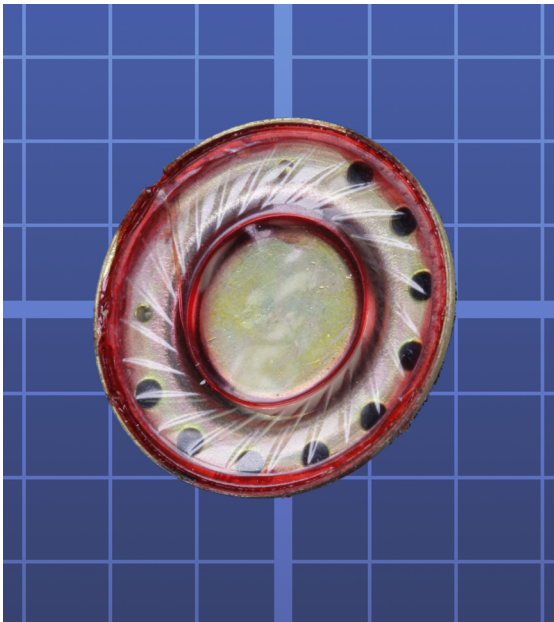


Figure 30-6. A miniature speaker measuring slightly more than 1/2" diameter and only 0.13" thick, designed for use in a cellular phone. It has an impedance of 150Ω.

Multiple Drivers

Generally speaking, a large-diameter speaker cone is more effective than a small cone at moving the greater volumes of air associated with reproduction of bass notes. However, the inertia of a large cone impairs its ability to vibrate at high frequencies.

To address this problem, a large speaker and a small speaker often share a single enclosure. A *crossover network* using coils and capacitors prevents low frequencies from reaching the small speaker and high frequencies from reaching the large speaker. The basic principle is shown in the simplified schematic in [Figure 30-7](#).

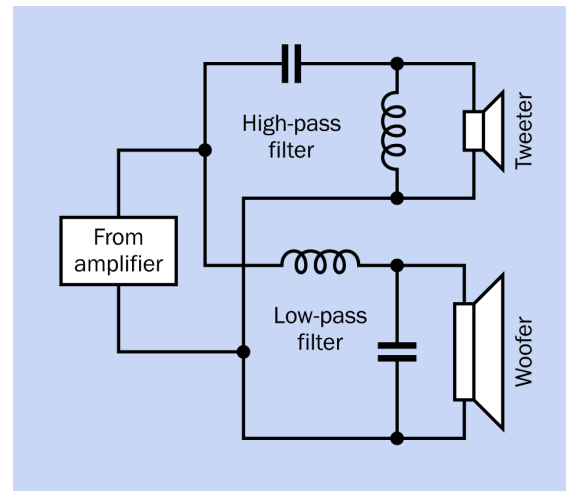


Figure 30-7. The basic principle of a crossover network.

Because the crossover network must be “tuned” to match the characteristics of the speakers, and because the combined sound pressure of the speakers must be relatively consistent over a wide range of frequencies, an actual network usually includes additional components.

Because the audio output from an amplifier consists of alternating current, polarized capacitors cannot be used. Polyester capacitors are common.

The small speaker in a pair is known as a *tweeter* while the large speaker is a *woofer*. Although

these appellations were whimsically coined, they have endured.

More than two speakers may be combined in an enclosure, in a wide variety of configurations.

Venting

A speaker radiates pressure waves from the back side of its cone as well as from the front, and because the waves from front and back are opposite in phase, they will tend to cancel each other out.

In a miniature speaker, this problem can be addressed simply by sealing the section of the enclosure at the rear. For larger components, a more efficient enclosure can be designed with a *vent* or *reflex port* at the front. Pressure waves from the back of the speaker are diverted over a sufficient distance inside the cabinet so that by the time they emerge through the port, they are approximately in phase with low frequencies from the front of the speaker, although the back wave will lag the front wave by one wavelength.

This design is referred to as a *bass-reflex enclosure*, and was almost universal in high-fidelity components until amplifiers became increasingly powerful during the 1960s. At that point, Acoustic Research, located in Massachusetts, marketed a product line in which speaker enclosures were sealed, the argument being that when an amplifier can deliver 100W per channel, efficiency is no longer an issue, and a sealed enclosure can eliminate compromises associated with a bass-reflex design.

Acoustic Research referred to their concept as “air suspension,” as the cushion of air in the sealed cavity helped to protect the speaker by limiting its excursion. This configuration is now often referred to as a *closed-box speaker*. Some audiophiles argue that it must always be inherently superior to a bass-reflex design, partly because of the one-wavelength lag time associated with a reflex port. However, as in many aspects of sound reproduction, the debate is inconclusive.

Resonance

The enclosure for a speaker will tend to have a dominant resonant frequency. This should be lower than the lowest frequency that the speaker will reproduce; otherwise, the resonance will emphasize some frequencies relative to others, creating unwanted peaks in the response.

One reason why high-quality speakers tend to be physically heavy is to reduce their resonant frequency. A modern Thiel speaker assembly, for instance, uses a front panel of particle board that is a full 2” thick. However, heavy enclosures are expensive to transport and inconvenient to locate or relocate in the home.

To address this problem, a tweeter and a woofer can be mounted in separate boxes. The enclosure for the tweeter can be very small, lightweight, and suitable for placement on a shelf, while the heavy box for the woofer can go on the floor. Human senses have difficulty locating the source of low-frequency sound, so the woofer can be located almost anywhere in a room. In fact, its single speaker can serve both stereo channels.

This configuration has become the default for computer speakers. It is also used in home-theater systems, where the woofer has now become a subwoofer capable of very low frequency reproduction.

Miniature Speakers

If an electronics project has an audio output, and the circuit board will be sharing an enclosure with a small speaker, the size of the box and the material from which it is fabricated will affect the sound quality significantly. A box made from thin hardwood may add resonance that sounds pleasing, if the speaker is being used just for simple electronic tones. By comparison, a metal box may sound “tinny.” A box fabricated from a plastic such as ABS will be relatively neutral, provided the plastic is reasonably thick (1/4” being preferable).

Variants

Electrostatic Speaker

The principle of an electrostatic speaker is the same as that of an electrostatic headphone. A charged membrane is stretched between two grids in front of it and behind it that act as electrodes. Because the membrane is so light, it responds with very little latency, and its large surface area creates a diffused sound that many audiophiles find pleasing. However, a high voltage is required to drive electrostatic speakers, and they are not cheap.

Powered Speakers

A unit containing its own driver electronics is referred to as a *powered speaker*, and is used almost universally with desktop computers, because the computer itself does not contain a power amplifier. Powered speakers may also enable a more versatile crossover network.

A subwoofer may have its own amplifier allowing control of the cutoff frequency above which the speaker will not attempt to reproduce sound. The electronics can include protection for a speaker against being overdriven.

Wireless Speakers

A wireless link between a stereo receiver and its speakers will eliminate the speaker wires that are normally necessary. However, the speakers themselves must be powered, and will have to be wired to electric outlets.

Innovative Designs

The need for small speakers in consumer products such as laptop computers has encouraged innovative designs. The speaker in Figure 30-8 is just 1" square, and its shape is easier to accommodate in a small product than the traditional circular speaker. In Figure 30-9, the interior of the same speaker shows that inductive coils are applied to a square plastic diaphragm.

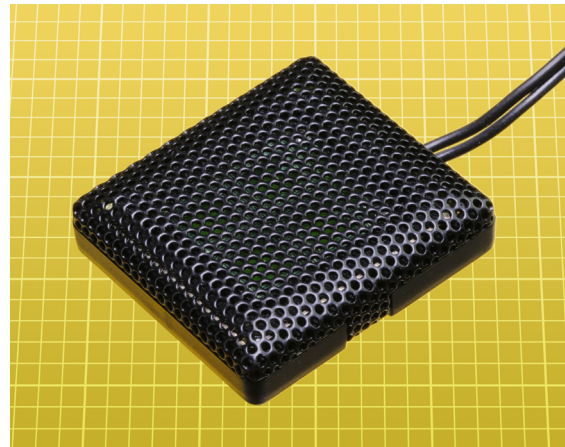


Figure 30-8. A speaker 1" square, suitable for use with a small electronic device.

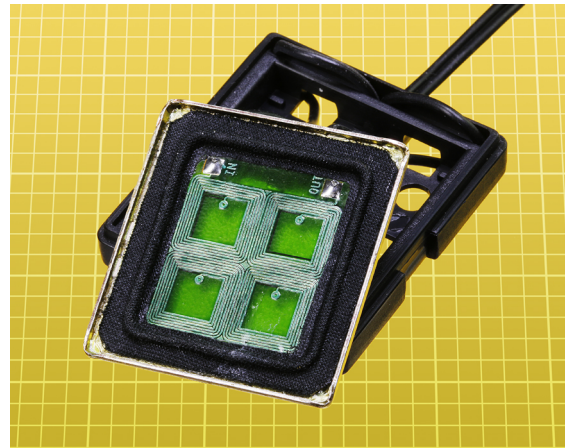


Figure 30-9. The speaker from the previous figure, opened to show its inductive coils applied to a square plastic diaphragm.

Values

The typical *impedance* for speakers in audio systems is 8Ω. Small speakers may have a higher impedance, which can be useful when driving them from devices that have limited power, such as a TTL-type 555 timer.

In the United States, the diameter of a circular speaker is usually expressed in inches. Speakers larger than 12" are rare for domestic use. A 4" speaker used to be considered minimal because

of its limited low-end frequency response, but much smaller speakers have become common in portable devices.

The low-end frequency response of a miniature loudspeaker designed to be surface-mounted on a circuit board will be very poor. The graph in [Figure 30-10](#) was derived from data supplied by the manufacturer.

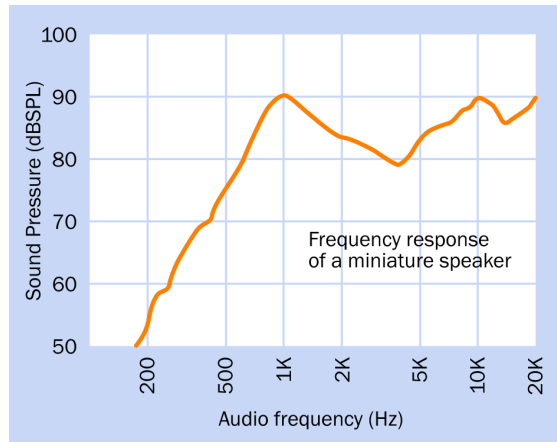


Figure 30-10. Frequency response for a speaker measuring 15mm x 15mm x 5mm. Small dimensions and lack of an enclosure result in a negligible low-end response.

Power rating for speakers is in watts, measured on a root-mean-square (RMS) basis.

Sensitivity is measured in decibels, at a distance of one meter, while a speaker is reproducing a single constant tone with a power input of 1W. A speaker assembly designed for undemanding home use may be rated at 85dB to 95dB.

Efficiency is a measure of sound power output divided by electrical power input. A value of 1% is typical.

What Can Go Wrong

Damage

As is the case with headphones, the most common problem affecting a speaker is damage caused by overdriving it. Because a low frequency requires larger excursions of a speaker cone to transmit an amount of energy comparable to that of a high frequency, loud bass notes can be hazardous to the cone. On the other hand, if an amplifier generates distortion (perhaps because it, too, is being overdriven), the harmonics created by the distortion can damage high-frequency speakers.

Magnetic Field

Even a small speaker (measuring 2" or less) contains a magnet that is sufficiently powerful to cause problems if it is in close proximity to other components, especially if sensors such as reed switches or Hall-effect switches are being used. Initial circuit testing should be done with the speaker as far away as possible, to eliminate it as a source of interference.

Vibration

Solder joints will be stressed if they are subjected to low-frequency vibration from a speaker. Loose parts may rattle, and parts that are bolted into place may become unbolted. The speaker itself may become loose. Thread-locker such as Loctite should be applied to nuts before they are tightened.

Index

Symbols

2N682 SCR, 2
4047B timer, 75
40TPS12 SCR, 2
4131-503 digital potentiometer, 62
4543B decoder, 168
4543B decoder chip, 168
555 timer, 70–73, 76, 79–84
5555 timer, 74
556 timer, 73
558 timer, 73
7447 decoder, 148
74HC00 logic gate, 93
74HC123 timer, 75
74HC163 counter, 132
74HC221 timer, 76
74HC4514 decoder, 147
74HC4515 decoder, 139, 147
74HC555 timer, 74
74HCT555 timer, 74
74LS123 timer, 75
74LS148 decoder, 139
74LS47 decoder, 148
7555 timer, 74

7556 timer, 74
BTA208X-1000B triac, 16
BTB04-600SL triac, 16
CAT5114 digital potentiometer, 64
CD4047B timer, 75
CMX60D10 relay, 27
D804 optocoupler, 37
DB3TG diac, 12
DC60S7 relay, 27, 28
HEF4528B timer, 75
LM339 comparator, 42
LM3914 driver, 236
LM741 op-amp, 49, 53
M74HC4538 timer, 75
MAC97A6 triac, 16
MAX6954 driver chip, 234
MC14538B timer, 75
NTE926 timer, 73
OPTEKD804 optocoupler, 34
PT-2040PQ transducer, 253
PUI XL453 audio indicator, 257
SSD1306 LED display module, 235
SSD1331 LED display module, 235

TIL311 driver, 236
WS0010 LED display module, 235
X0403DF SCR, 2, 5

A

A-weighting audio standard, 252
absolute zero, 173
acoustically transparent headphones, 264
activator, electroluminescent, 244
active matrix LCD, 161
additive primaries, 166
address of tap, digital potentiometer, 62, 63
Adobe 1998, 167
alert, 249, 257
alternating current, 172
American Standard Code for Information Interchange (ASCII), 163
AMOLED (active-matrix OLED), 246
analog input/output, 66

- analog integrated circuits
 - comparator, 39–48
 - (see also comparator)
 - digital potentiometer, 59–67
 - (see also digital potentiometer)
 - op-amp, 49–58
 - (see also op-amp)
 - optocoupler, 33–37
 - (see also optocoupler)
 - solid-state relay, 25–31
 - (see also solid-state relay)
 - timer, 69–87
 - (see also timer)
- analog multiplexer, 151, 152
- analog output optocoupler, 36
- analog-digital converter, 39, 47, 59
- AND gate, 45, 90, 91, 92, 97, 98, 100, 102, 103
- anode
 - LED display, 232
 - LED indicator, 213
 - vacuum-fluorescent display (VFD), 240
- arithmetical operations, binary, 91
- arithmetical operators, 89
- ASCII (see American Standard Code for Information Interchange)
- astable multivibrator, 46, 55, 69
 - (see also timer)
- astable timer, 70, 72, 77, 80
- asynchronous counter, 118, 134
- asynchronous flip-flop, 107
- asynchronous input, 123
- audio alerts
 - audio indicator, 257–260
 - transducer, 249–255
- audio frequency, 54, 251, 252, 253, 254, 257, 258
- audio indicator, 249, 257–260
 - PUI XL453, 257
 - current, 260
 - design, 257–258
 - duty cycle, 260
 - formats, 258
 - frequency, 258, 260

- function, 257
- history, 258
- potential problems, 260
- sound intensity, 260
- sound patterns, 258
- use of, 260
- values, 259
- variants, 258
- voltage, 259
- volume control, 260
- wiring, 260

audio transducer (see transducer)

B

- back-EMF, 29, 30, 87
- backlighting, electroluminescent, 167
- backplane, 161
- ballast, 180, 192, 194, 225
- band gap, 207
- bandpass filter, 59
- bandwidth, 36
- bass-reflex enclosure, 270
- bayonet base, 177
- BCD output, 135
- beeper, 249, 257
- bi-pin tube, 194
- bidirectional multiplexer, 151
- bidirectional optocoupler, 36
- bidirectional shift, 125
- bidirectional thyristor, 11
- binary addition, 91
- binary arithmetic, 89, 92
- binary counter, 134
- binary digit, 121, 134
- binary ripple counter, 121
- binary-coded decimal (BCD) output, 135
- bipolar transistor, 1, 18, 70
- bistable displays, 167
- bistable multivibrator, 46, 69, 107
 - (see also flip-flop)
- bit, 121, 134
- bitmap, 159, 227
- black body radiation, 173
- Boolean algebra, 89
- Boolean operators, 89, 91

- breakdown voltage, 1, 4, 20, 184
- breakover frequency, 53
- breakover level, 12
- breakover voltage, 4, 14, 15, 20
- bubble, 90, 116, 123, 133, 155
- buffer, 90, 128
- burn in, 160
- burst mode, 84
- buzzer, 249, 257, 258

C

- Camenzind, Hans, 70
- canalphones, 264
- candela, 178
- candlepower, 178
- capacitor
 - coupling, 84
 - light-emitting, 244
- carbon arc, 171
- cascade, 134
- cascaded encoders, 142
- cascading, 157
- cathode, 188, 192, 232, 239
- cathode vacuum-fluorescent display (VFD), 240
- cathode-ray tube (CRT), 160, 239, 244
- CCFL (see cold-cathode fluorescent lamp)
- CCT (see correlated color temperature)
- cd (candela), 178
- ceramic wafer, 250, 258
- CFL (see compact fluorescent lamp)
- channel current, 155
- channels, 42, 52, 153
- charge storage, 23, 24
- chemiluminescence, 244
- chip select pin (CS), 63, 64
- cilia, 261
- circumaural headphones, 264
- clipped output, 50
- clock, 121
- clock input, 131, 133
- clocked comparator, 47
- clocked in, 121

closed-back headphones, 264
 closed-box speaker, 270
 CLR operation, 133
 CMOS 555 timer, 74
 CMOS chips, 48, 94
 CMYK system, 166
 CO₂ lasers, 201
 code hopping encoder, 139
 coherent light, 197, 200
 cold-cathode fluorescent lamp (CCFL), 191, 194
 cold-cathode fluorescent panel, 167
 color
 fluorescent lights, 194, 241, 242
 in LCDs, 166–167
 of LEDs, 205, 207, 209–211, 212, 219–222, 222, 226
 primary, 166–167
 color rendering index (CRI), 174
 color super twisted nematic (CSTN) LCD, 161
 color temperature, 173, 209, 219, 222
 color-rendering index (CRI), 212
 common anode, 232
 common cathode, 232
 common mode voltage range, 53
 common pin, 161
 commutating dv/dt, 23
 compact fluorescent lamp (CFL), 191
 comparator, 39–48, 70, 83, 90
 LM339, 42
 analog-digital converter, 47
 AND gate, 45
 bistable multivibrator, 46
 clocked, 47
 continuous converter, 47
 design, 39–42
 digital, 39
 dual, 42
 function, 39
 hysteresis, 39, 44
 latched, 47
 level shifter, 46
 and op-amps, 42, 52
 potential problems, 47–48

relaxation oscillator, 46
 use of, 44–47
 values, 43–44
 variants, 42–43
 window, 42
 window comparator, 46
 zero point finder, 47
 conduction angle, 7, 9
 cone, 268
 confused inputs, 48
 continuous converter, 47
 converter, serial to parallel, 107
 correlated color temperature, 222
 counter, 107, 121, 131–138
 74HC163, 132
 asynchronous, 118, 134
 binary, 134
 binary ripple, 121
 clock sources, 135
 decade, 135
 decoded output, 134
 descending output, 136
 design, 132–134
 divide-by-16, 134
 function, 131–132
 glitches, 138
 gray code, 131
 hexadecimal, 134
 modulus-16, 134
 modulus/modulo, 132
 multiple stages, 136
 octal, 134
 output states, 136
 pin identifier, 133
 potential problems, 137–138
 programmable, 137
 ring, 131, 134
 ripple, 107, 134, 148
 rising edge and falling edge, 136
 schematic representation, 131–132
 single and dual, 136
 synchronous, 134, 136
 values, 137
 variants, 134–137
 coupler, output, 198
 coupling capacitor, 84
 CRI (see color rendering index)

crossover network, 269
 crosstalk, 161
 crowbar overvoltage protection, 2
 CRT (see cathode-ray tube)
 crystal, 249
 crystal lasers, 201
 crystal-set radio, 265
 CS (chip select pin), 63, 64
 current
 alternating, 172
 audio-indicator, 260
 channel, 155
 direct, 172
 forward, 211
 gate threshold, 18
 holding, 2, 18
 latching, 2, 18
 leakage, 156
 transducer, 254
 current amplifier, 2, 18
 Current Transfer Ration (CTR), 36

D

D-type flip-flop, 114, 116
 data distributor, 140, 146, 152
 data selector, 140, 146, 151, 152
 (see also multiplexer)
 data transfer, 62
 Db (see decibels)
 dBA, 253
 debounced, 66
 decade counter, 135
 decibel, 251–252, 259, 265
 decoded output counter, 134
 decoder, 134, 139, 140, 143, 145–149, 151, 153, 157
 7447, 148
 74HC4514, 147
 74HC4515, 147
 74LS47, 148
 design, 148
 function, 145–146
 input devices, 145–146
 LED driver, 146
 potential problems, 149–149
 schematic symbol, 146–146

- seven-segment, 146
- similar devices, 146
- use of, 149
- values, 148
- variants, 148
- decoder chip, 151
- demultiplexer, 140, 146, 151
- derated relay, 30
- derating curve, 213
- descending output counter, 136
- diac, 1, 11–14, 15, 22
 - DB3TG, 12
 - abbreviations, 14
 - design, 12–13
 - function, 11
 - potential problems, 14
 - switching AC, 13
 - symbol variants, 11
 - values, 14
 - variants, 13
- diaphragm, 262, 268
- dies, 219
- differential multiplexer, 152
- digipot (see digital potentiometer)
- digital comparator, 39
- digital integrated circuits
 - counter, 131–138
 - decoder, 145–149
 - encoder, 139–143
 - flip-flop, 107–119
 - logic gate, 89–106
 - multiplexer, 151–157
 - shift register, 121–129
- digital multiplexer, 151
- digital potentiometer, 59–67, 156
 - 4131-503, 62
 - CAT5114, 64
 - address of tap, 62, 63
 - advantages/disadvantages, 59–60
 - connections and modes, 64
 - data transfer, 62
 - design, 60–61
 - function, 59–60
 - higher resolution, 66
 - I2C protocol, 63
 - potential problems, 66
 - programmable, 59

- SPI protocol, 62
- taper, 62
- up/down protocol, 63
- use of, 66
- values, 65–66
- variants, 61–65
- volatile/nonvolatile memory, 61
- digital-analog converter, 39, 47, 59
- digitally adjustable potentiometer (see digital potentiometer)
- digitally controlled potentiometer (see digital potentiometer)
- digitally programmed potentiometer (DPP) (see digital potentiometer)
- digipot (see digital potentiometer)
- diode, PIN, 35
- diode, protection, 30
- direct current, 172
- disallowed state, 135
- discrete semiconductor
 - diac, 22
- discrete semiconductors, 1
 - (see also thyristor)
- dispersion, 178, 208, 225
- display
 - electroluminescent (see electroluminescent (EL) display)
 - LED (see LED display)
- display module, 164, 231
- divide-by-16 counter, 134
- dot-matrix displays, 163–165, 230, 234–236, 239
- double super twisted nematic LCD, 161
- drift, 48
- driver, 267
- driver chip, 233–234
- driver, dynamic, 262
- dual comparator, 42
- dual counter, 136
- dual digital potentiometer, 61
- dual inputs, 50–51, 127
- dual monostable timer, 75–76, 77
- dual voltage device, 39
- dual voltage power supply, 50

- duty cycle, 260
- dv/dt triggering, 8
- dynamic driver, 262
- dynamic transducer, 262

E

- e-ink displays, 167
- ear canal headphones, 264
- earbuds, 261, 263, 264
- earphone, 261, 264
- edge pass, 69
- edge triggering, 69
- edge-triggered shift register, 123
- efficiency, 272
- EL (see electroluminescent (EL) display)
- electret, 264
- electric newspaper, 236
- electroluminescence, 243
- electroluminescent (EL) display, 243–247
 - derivation, 244
 - design, 243
 - flexible ribbons, 245
 - function, 243
 - OLED, 246–247
 - panels, 244–245
 - phosphors, 244
 - rope light, 245
 - variants, 244–247
- electroluminescent backlighting, 167
- electromagnetic relay, 33
- electromagnetic transducer, 250, 253
- electronic ballast, 193, 225
- electronic paper displays, 167
- electrostatic headphones, 263
- electrostatic speaker, 271
- Enable, 116
- encoded output, 134
- encoder, 139–143, 152
 - 74HC4515 decoder, 139
 - 74LS148 decoder, 139
- cascaded, 142
- code hopping, 139
- design, 140–141

- function, 139–140
- potential problems, 143
- priority, 140, 141
- rotary, 139
- rotational, 66
- schematic symbol, 139
- similar devices, 140
- simple, 141
- tri-state, 141
- use of, 142
- values, 142
- variants, 141

equal-loudness contours, 252

erratic output, 48

F

- falling voltage, 43
- falling-edge triggered shift register, 123
- fanout, 94
- feedback
 - negative, 51–52
 - positive, 44
- ferromagnetic disc, 250
- FF (see flip-flop)
- fiber lasers, 201
- FIFO (see first-in, first-out (FIFO) shift register)
- filament, 171, 239
- film-compensated super twisted nematic (FSTN) LCD, 161
- first-in, first-out (FIFO) shift register, 123
- flange base, 177
- flexible ribbon, 245
- flicker, 193
- flip-flop, 46, 70, 89, 107–119, 121, 132
 - asynchronous/synchronous, 107
 - D-type, 114
 - design, 107–116
 - forbidden states, 110–112
 - function, 107
 - JK, 112–113
 - master-slave, 113
 - NAND-based, 108–109

- NOR-based, 109–110
- packaging, 117
- potential problems, 118–119
- summary of types, 116
- use of, 118
- values, 117–118
- variants, 116–117

flip-flop emulation, 82

floating inputs, 127

floating pins, 105

fluoresce, 192

fluorescence, 191, 239

fluorescent lamp, 191

fluorescent light, 24, 167, 174, 191–196, 207, 217, 223, 239
(see also vacuum-fluorescent display)

ballast and starter, 192

brightness, 195

CCFLs, 194

comparisons, 194

design, 191–195

flicker, 193

function, 191

potential problems, 195

sizes, 194–194

spectrum, 195

tubes, 191

values, 195

variants, 193

fluorescent tubes, 191

flux, 177

foot-candles, 178

forbidden states, 110–112

frequency

- audio, 54, 251, 252, 253, 254, 257, 258, 260
- breakover of op amp, 53
- unity gain, 53

frequency response, 265

fused quartz, 176

G

- gain, 50
- gain medium, 198
- game sound, 85
- gamut, 167, 207

- ganged analog potentiometer, 61
- gas-discharge device, 191
- gate threshold current, 18
- gate turn-off thyristor (GTO), 1, 11, 15
- gated circuit, 113
- glitches, 138, 149
- glow discharge, 185
- gray code counters, 131
- ground loops, 35
- GTO (see gate turn-off thyristor)

H

- half adder, 92
- halogen lighting, 176, 217, 221
- harmonics, 266
- headphone, 261–266
 - acoustically transparent, 264
 - balanced armature, 264
 - circumaural, 264
 - closed-back, 264
 - design, 261–262
 - distortion, 266
 - electret, 264
 - electrostatic, 263
 - frequency response, 265–266
 - function, 261
 - impedance, 266
 - in-ear, 264
 - intensity, 265
 - mechanical design, 264–265
 - moving coil, 262–263
 - noise-cancelling, 264
 - open-back, 264
 - potential problems, 266
 - supra-aural, 264
 - values, 265–266
 - variants, 262–265
- headset, 264
- hearing aids, 261
- heat sensitivity, 14
- heat sink, 28, 206, 222
- Hertz, Heinrich Rudolph, 262
- hexadecimal counter, 134
- high linearity optocouplers, 36
- high logic state, 89
- high speed optocoupler, 36

high-brightness LED, 205, 217, 227
high-intensity LED, 217
high-output LED, 217
high-pass filter, 55
high-power LEDs, 217
Hold state, 109, 114
holding current, 2, 18
holding level, 12
hysteresis, 42, 44–45, 48, 83, 185
hysteresis range, 43
hysteresis zone, 39

I

I2C protocol, 62, 63
IEM (see in-ear monitor)
ignition voltage, 184
illuminance, 178
impedance, 271
in-ear earphones, 264
in-ear headphones, 264
in-ear monitor, 264
incandescence, 191, 243
incandescent bulbs, 219
incandescent lamp, 11, 15, 24, 171–181, 200, 223
 advantages/disadvantages, 179–180
 base variants, 177
 derating, 180
 design, 172–175
 efficacy, 179
 efficiency, 179
 function, 171–172
 halogen or quartz halogen, 176
 illuminance, 178
 intensity, 178
 mean spherical candlepower (MSCP), 178
 miniature lamps, 175–176
 non-incandescent sources, 174
 oven lamp, 176
 panel-mount indicator lamp, 176
 potential problems, 180

power, 177–178
power consumption, 175
schematic symbols, 171
spectrum, 173
use of, 179–180
values, 177–179
variants, 175–177
incandescent lamps, 206, 217
incoherent light, 159
increment/decrement protocol, 62
Indiglo electroluminescent displays, 244
inductor, 193
infrared emitters, 211
inhibit pin, 153
input
 analog, 66
 asynchronous, 123
 clock, 131, 133
 confused, 48, 58
 dual, 50–51, 127
 floating, 127
 inverting/noninverting, 39, 51
 parallel, 123, 134
input bias current, 53
input devices, decoder, 145
input differential voltage range, 53
input offset voltage, 43, 53
input voltage, 155
instant-start ballast, 194
integrated circuits (see analog integrated circuits, digital integrated circuits)
inter-integrated circuit, 62
internal sensors, 35
internal series resistor, 210
intrinsic layer, optocoupler, 35
invalid number, 135
inversion, logic gates, 90
inverter, 36, 89, 91, 114
inverting input, 39, 51
ionization, 184
isolation transformers, 34

J

jam loaded, 134
jam-type flip-flop, 108
jam-type parallel data input, 124
jitter, 44
JK flip-flop, 112–113, 116

K

keyboard, polling, 127

L

ladder, 60
lamp
 oven, 176
 panel-mount indicator lamp, 176
lamp lenses, 176
laser, 197–203
 CO2 lasers, 201
 coherent light, 200–201
 common applications, 202
 crystal lasers, 201
 design, 198–201
 fiber lasers, 201
 function, 197
 laser diode, 197, 198–199
 potential problems, 202
 use of, 202
 values, 201
 variants, 201
laser diode, 197, 198–199
latch, 107, 112
latch function, 43
latched comparator, 47
latching current, 2, 18
latching relay, 4
LCD (liquid-crystal display), 159–170, 207, 227, 246
 4543B decoder, 168
 active and passive types, 161
 additional segments, 162–163
 alphanumeric display module, 168–169
 backlighting, 167

- color, 166–167
- color super twisted nematic (CSTN), 161
- crystal types, 161
- design, 159–160
- dot-matrix displays, 163–165
- double super twisted nematic, 161
- film-compensated super twisted nematic (FSTN), 161
- function, 159
- LED comparisons, 228
- numeric display module, 167–168
- potential problems, 169–170
- reflective, 161
- seven-segment displays, 161–162
- super twisted nematic (STN), 161
- transreflective, 161
- twisted nematic (TN), 161
- use of, 167–169
- variants, 160–167
- zero-power displays, 167
- LCD display (see LCD (liquid crystal display))
- leakage
 - capacitors, 77
 - diac, 12
 - SCR, 2
 - solid-state relay (SSR), 26
 - triac, 17
- leakage current, 156
- LED (generic), 26, 159, 167, 174, 227, 243
- LED area lighting, 24, 191, 195, 217–226, 227
 - color variation, 222
 - comparisons, 220–222, 223–224
 - costs and efficiency, 218
 - design, 218–223
 - dimming, 222
 - efficacy, 222
 - function, 217–218
 - heat dissipation, 222
 - high-brightness, 217
 - high-output, 217
 - high-power, 217
 - potential problems, 225
 - schematic symbol, 218
 - ultraviolet output, 222
 - values, 225
 - variants, 223–224
 - visible differences, 220–220
- LED display, 146, 227–237
 - LM3914, 236
 - MAX6954, 234
 - SSD1306, 235
 - SSD1331, 235
 - TIL311, 236
 - WS0010, 235
 - additional segments, 229–230
 - design, 228
 - dot-matrix displays, 230, 234–236
 - driver chips and multiplexing, 233–234
 - function, 227–228
 - LCD comparisons, 228
 - multiple bar display, 232, 236
 - multiple numerals, 229
 - one-digit hexadecimal dot matrix, 236
 - pixel arrays, 231, 235–236
 - potential problems, 237
 - seven segment numeral, 228
 - seven-segment displays, 228–229, 232–233
 - single light bar, 232
 - sixteen-segment driver chip, 234
 - use of, 232–236
 - values, 232
 - variants, 228–232
- LED indicator, 205–215, 218, 227
 - color rendering index (CRI), 212
 - common usage, 206
 - comparisons to other light emitters, 214
 - design, 207
 - diffusion, 209
 - efficacy, 208
 - forward current, 211
 - forward voltage, 212
 - function, 205–207
 - high-brightness, 205
 - infrared, 211
 - intensity, 208
 - internal resistor, 210
 - LEDs in parallel, 214
 - life expectancy, 212
 - light output and heat, 212
 - low-current LEDs, 211
 - multicolor LEDs and color mixing, 207
 - multicolored, 210
 - multiple series LEDs, 214
 - other applications, 215
 - polarity, 213
 - potential problems, 215
 - schematic symbols, 206
 - series resistor value, 214
 - size and shape, 208
 - ultraviolet, 211
 - use of, 213–215
 - values, 211–213
 - variants, 208–211
 - view angle, 213
 - wavelength and color temperature, 209–210
- lens, 183
- level shifter, 46
- light
 - coherent, 197, 200–201
 - incoherent, 159
 - monochromatic, 197
- light wires, 243
- light-emitting capacitor, 244
- light-emitting diode (see LED)
- linear relationship, 50
- linear taper, 62
- linearity versus saturation, 42
- liquid-crystal display (see LCD)
- logarithmic taper, 62
- logic chip, 139
- logic gate, 89–106, 108, 126, 140
 - 74HC00, 93
 - 4000 series, 94
 - 7400 series, 93
 - binary arithmetical operations, 91
 - Boolean notation, 91
 - chip families, 95–96
 - design, 89

- function, 89
- gates and inputs, 90, 96–103
- inversion, 90
- part numbers, 94–95
- potential problems, 105–106
- uses, 103
- variants, 93
- logic state, 89, 107
- logic, positive, 90
- logic, transistor-transistor, 70, 93
- logic-output optocoupler, 36
- loudspeaker, 84, 249, 263, 267
- low logic state, 89
- low-current LEDs, 211
- low-pass filter, 54
- Lower State Transition Voltage (LSTV), 43
- lumen (lm), 177, 195, 218, 225
- lumens per watt, 179, 195
- luminaire, 217
- luminary, 217
- luminescence, 191, 243
- luminous flux, 177, 195, 208, 218
- lux, 178

M

- magnetic ballast, 225
- maintaining voltage, 185
- maser, 197
- master-slave flip-flop, 113, 116
- mc (millicandela), 178
- mean spherical candlepower (MSCP), 176, 178
- metal-oxide semiconductor field-effect transistor (see MOSFET)
- metastability, in flip-flops, 119
- microcontroller, 29, 59
- microphone, 262
- millicandelas, 187, 208
- miniature speakers, 270
- mismatching, 266
- mixed-signal device, 59
- MOD, 133
- mode select pins, 125
- modulo/modulus, 132
- modulus-16 counter, 134
- monochromatic light, 197
- monostable multivibrator, 69
 - (see also timer)
- monostable timer, 69, 71, 77, 79
- MOSFET (metal-oxide semiconductor field-effect transistor), 26, 70
- motion detector, passive infrared, 211
- motor, AC, 11
- moving coil headphones, 262
- MSCP (see mean spherical candlepower)
- multicolor LED, 207
- multiple bar display, 232, 236
- multiple series LED, 214
- multiple-stage counter, 136
- multiplexer, 28, 140, 146, 151–157
 - analog, 151, 152
 - bidirectional, 151
 - demultiplexer, 140, 146, 151, 152, 157
 - design, 153
 - differential, 152
 - digital versus analog, 151
 - function, 151
 - pin identifiers, 154–155
 - potential problems, 157
 - schematic symbol, 153
 - similar devices, 152–153
 - use of, 156
 - values, 155
 - variants, 155
- multiplexing, 168, 233–234, 241
- multisource lighting
 - electroluminescent (EL) display, 243–247
 - LED display, 227–237
 - vacuum-fluorescent display (VFD), 239–242
- multivibrator, 69
 - (see also timer)
 - astable, 46, 55, 69
 - bistable, 46, 69, 107
- MUX (see multiplexer)

N

- NAND gate, 37, 90, 91, 93, 94, 95, 97, 98, 100, 102, 103, 108, 109, 123, 127, 135
 - (see also NAND-based SR flip-flop)
- NAND-based SR flip-flop, 108–109, 116
- nanometer, 195, 209
- negative feedback, 49, 51–52
- negative resistance, 185
- neon bulb, 31, 183–190, 191, 193
 - construction, 184
 - efficiency, 187
 - function, 183
 - how it works, 183–186
 - ionization, 184
 - life expectancy, 188
 - light output, 187
 - negative resistance, 185–186
 - Nixie tubes, 189
 - potential problems, 189
 - power-supply testing, 188
 - ruggedness, 187
 - use of, 186–189
 - variants, 189
- neon bulbs, 206
- neon indicator, 183
- neon lamp, 183
- neon lamp assembly, 183
- neutral value, 49
 - (see also zero value)
- Newtons per square meter (Pa), 251
- Nixie tubes, 189
- noise-cancelling headphones, 264
- noise-creating device, 249, 267
- noninverting input, 39, 51
- nonvolatile memory, 61
- NOR gate, 90, 91, 97, 98, 100, 102, 103, 108, 109
 - (see also NOR-based SR flip-flop)
- NOR-based SR flip-flop, 109–110
- numeric display module, 167–168

NV (see nonvolatile memory)

O

octal counter, 134
OE (output-enable), 126
offset null adjustment, 56
OLED (organic light-emitting diode), 217, 227, 243, 246–247
on-resistance, 156
one-digit hexadecimal dot matrix, 236
op-amp, 39, 49–58, 90
 LM741, 49, 53
 calculating amplification, 53
 and comparators, 52
 confused input, 58
 controlling the gain, 53
 design, 49–52
 differences from comparator, 42
 dual inputs, 50–51
 as high-pass filter, 55
 as low-pass filter, 54
 negative feedback, 51–52
 offset null adjustment, 56
 oscillating output, 57
 potential problems, 57–58
 as relaxation oscillator, 55
 as single power source, 56
 use of, 53–57
 values, 52
 variants, 52
 what it does, 49
open collector, 40
open loop mode, 51
open loop operation, 42
open loop voltage gain, 53
open-back headphones, 264
open-drain outputs, 126
operating voltage, 259
operational amplifier (see op-amp)
optical maser, 197
optical switch, 34
optocoupler, 25, 33–37, 215
 D804, 37
 OPTEKD804, 34
 analog output, 36
 basic types, 36
 bidirectional, 36
 design, 34–34
 function, 33–34
 high speed, 36
 high-linearity, 36
 internal sensors, 35–36
 logic-output, 36
 potential problems, 37
 schematic symbols, 34
 use of, 36–37
 values, 36
 variants, 34–36
OR gate, 90, 91, 97, 98, 99, 100, 102, 103, 140, 142
organic LED (see OLED (organic light-emitting diode))
oscillating output, 47, 57
oscillator, relaxation, 12, 46, 55
output
 analog, 66
 clipped, 50
 descending, 136
 encoded, 134
 erratic, 48
 open-drain, 126
 oscillating, 47, 57
 parallel, 123
 push-pull, 42
 three-state, 142
 types of, 134
 ultraviolet, 222
 weighted, 134
output coupler, 198
output mode, comparator, 42
oven lamp, 176
overvoltage protection, 2, 6, 155

P

Pa (Pascals), 251
panel, 243
Panescent electroluminescent lighting, 244
parallel input, 134
parallel-enable pin, 134
parallel-in, parallel-out (PIPO) shift converters, 125
parallel-in, serial-out (PISO) shift register, 124
parallel-serial converters, 124
parchmentized thread, 172
passive infrared motion detector, 211
passive matrix LCD, 161
phase angle, 7, 13
phase control, 2, 7–7, 13, 22
phones (see headphone)
phosphor, 192, 239, 244
phosphorescence, 244
photocell, 35
photodarlington, 34
photodiode, 26, 34
photon, 207
photoresistor, 34
phototransistor, 26, 34, 45
photovoltaic PIN diode, 35
piezein, 250
piezo, 258
piezoelectric alerts (see audio indicator)
piezoelectric transducer, 249, 250, 253
pin
 common, 161
 floating, 105
 inhibit, 153
 parallel-enable, 134
 reset, 70
 trigger, 69
pin base, 177
PIN diode, 35
pin identifier, 133
pip, 184
PIPO (see parallel-in, parallel-out (PIPO) shift converters)
pixel arrays, 231, 235–236
PMOLED (passive-matrix OLED), 246
PN junction, 207
PNPN device, 3
polling a keyboard, 127
positive feedback, 44
positive logic, 90

- potentiometer, 35
 - (see also digital potentiometer)
- power rating, 272
- powered speaker, 271
- primary colors, 166–166
- priority encoder, 140, 141
- programmable counter, 137
- programmable digital potentiometer, 59
- programmable unijunction transistor (PUT), 2, 9
- propagation delay, 44, 134
- protection diode, 30
- pullup resistor, 36, 40, 48, 71, 126, 153
- pulse-width modulation, 7, 194, 211
- pumping a laser, 198
- push-pull output, 42
- pushbutton, 66, 71
- pushbutton protocol, 62

Q

- quad digital potentiometer, 61
- quadrants, 17–18
- quartz crystal, 131, 135

R

- race condition, 112
- radiant flux, 177
- radiant luminous efficacy (LER), 179, 187, 208, 222
- radiant luminous efficiency (LFR), 179
- radiation, black body, 173
- rail-to-rail values, 42
- rapid-start ballast, 194
- rated voltage, 259
- Rayleigh scattering, 202
- RC network, 70
- rectifier, 2
- reference voltage, 70
- reflective display (see LCD (liquid crystal display))

- reflective LCD, 161
- reflective primaries, 166
- reflex port, 270
- regenerative device, 2
- register, 121
- relaxation oscillator, 12, 46, 55
- relay
 - electromagnetic, 33
 - latching, 4
 - solid-state (see solid-state relay)
 - zero-crossing, 25, 27
- remainder, 132
- reproducers
 - headphone, 261–266
 - speaker, 267–272
- reset pin, 70
- reset state, 109
- resistor array, 232
- resolution, 43
- resonant cavity, 198
- response time, comparator, 42
- restricted combination, 112
- retriggering, timer, 69
- RGB LCD monitors, 167
- RGB primaries, 166
- rheostat mode, 64
- ring counter, 121, 131, 134
- ringing, 112
- ripple counter, 107, 134, 148
- rising voltage, 43
- rising-edge triggered shift register, 123
- rope light, 243, 245
- rotary encoder, 139
- rotary switch, 151
- rotational coder, 139
- rotational encoder, 66

S

- saturation versus linearity, 42
- SCR (silicon-controlled rectifier), 1–9, 11, 15, 26, 34
 - 2N682, 2
 - 40TPS12, 2
 - X0403DF, 2, 5
 - abbreviations, 6

- AC current applications, 5
- breakdown and breakover voltage, 4
- concept demo, 4
- design, 1–2
- function, 1
- internal configuration, 3–5
- overvoltage protection, 7
- phase control, 7–7
- potential problems, 8–9
- switching behavior, 2
- use, 6–8
- values, 5
- variants, 5, 6
- screw-in lamps, 177
- SCS (see silicon-controlled switch)
- segments, 239
- self-drive transducer circuit, 254
- sensitivity, 272
- sensor, 34, 35, 249, 262
- serial peripheral interface (see SPI)
- serial-in, parallel-out (SIPO) shift register, 123, 124
- serial-in, serial-out (SISO) shift register, 123, 124
- serial-parallel converter, 123
- serial-to-parallel converters, 107
- series resistor, 207
- series resistor value, 214
- settling time, 149, 157
- setup time, 125
- seven-segment decoder, 146
- seven-segment displays, 228–229, 232–233
- seven-segment numeral, 228
- shift register, 63, 121–129, 131, 146
 - abbreviations and acronyms, 123
 - arithmetical operations, 127
 - buffering, 128
 - design, 122
 - function, 121–122
 - parallel outputs and inputs, 123–124
 - polling the keyboard, 127
 - potential problems, 128–129

- power considerations, 126
- preloading, 127
- schematic representation, 122
- three-state output, 126
- use of, 126–128
- values, 125–126
- variants, 124–125
- shift registers, 107
- shortened modulus, 135
- sidac, 14
- silicon-controlled rectifier (see SCR)
- silicon-controlled switch (SCS), 1, 11, 15
- simple encoder, 141
- single counter, 136
- single light bar, 232
- single power source, 56
- single-inline package (SIP), 27
- single-input gates, 90
- single-source lighting
 - fluorescent light, 191–196
 - incandescent lamp, 171–181
 - lasers, 197–203
 - LED area lighting, 217–226
 - LED indicator, 205–215
 - neon bulb, 183–190
- sink current, 44
- SIPO (see serial-in, parallel-out (SIPO) shift register)
- siren, 258
- SISO (see serial-in, serial-out (SISO) shift register)
- sixteen-segment driver chip, 234
- slew rate at unity gain, 53
- snubber, 30
- snubber network, 23
- snubberless triac, 23
- solid-state analog switch, 28
- solid-state relay (SSR), 1, 11, 15, 25–31, 33, 36, 215
 - (see also optocoupler)
 - CMX60D10, 27
 - DC60S7, 27, 28
 - advantages/disadvantages, 25
 - design, 26–27
 - function, 25–26
 - instantaneous versus zero crossing, 27

- NC and NO modes, 28
- packaging, 28
- potential problems, 30
- solid-state analog switch, 28
- use of, 29–30
- values, 29
- variants, 27–28
- sound pressure level, 251–252, 251
- sound reproduction device, 249, 267
- sound sources (see audio alerts, reproducers)
- spatial coherence, 200
- spatial distribution, 213
- speaker, 249, 267–272
 - closed-box, 270
 - construction, 268
 - design, 267–270
 - electrostatic, 271
 - function, 267
 - innovative designs, 271
 - miniature speakers, 270
 - multiple drivers, 269–270
 - potential problems, 272
 - powered, 271
 - resonance, 270
 - values, 271
 - variants, 270–271
 - venting, 270
 - wireless, 271
- spectral lines, 174
- SPI (serial peripheral interface), 62
- spontaneous emission, 198
- SPST switch, 25
- sputtering, 188
- SR flip-flop (see NAND-based SR flip-flop)
- sRGB, 167
- SSR (see solid-state relay)
- standard LED, 205, 227
- starter, 192, 225
- starting voltage, 184
- state diagram, 135
- steradians, 178
- stimulated emission, 198
- striking voltage, 184
- strip lights, 223

- strobe, 153
- super twisted nematic (STN) LCD, 161
- supra-aural headphones, 264
- surface-mount transducers, 251
- swapped voltages, 48
- switch, 25, 71, 107
 - rotary, 151
 - silicon-controlled (SCS), 1, 11, 15
 - solid-state analog switch, 28
 - SPST, 25
- switch bounce, 109
- synchronous circuit, 113
- synchronous counter, 134, 136
- synchronous flip-flop, 107

T

- T-type flip-flop, 116
- tap, 60
- tap address, 62, 63
- taper, 62
- task lighting, 217
- TFT (see thin-film transistor)
- THD (see total harmonic distortion)
- thermal compound, 30
- thick phosphor, 244
- thin-film OLED, 243
- thin-film transistor (TFT), 161, 246
- 3-wire programmable potentiometers, 63
- three-state output, 126, 142
- three-wire chips, 64
- threshold voltage, 207
- through hole, 206
- thyatron, 1
- thyristor
 - diac, 11–14, 15
 - SCR, 1–9
 - triac, 15–24
- timer, 69–87
 - 4047B, 75
 - 555, 70–85
 - 5555, 74
 - 556, 73
 - 558, 73

- 74HC123, 75
- 74HC221, 76
- 74HC5555, 74
- 74HCT5555, 74
- 7555, 74
- 7556, 74
- CD4047B, 75
- HEF4528B, 75
- M74HC4538, 75
- MC14538B, 75
- NTE926, 73
- astable mode, 70, 72, 77, 80
- bipolar version, 70
- burst mode, 84
- CMOS 555, 74
- CMOS version, 71
- and coupling capacitors, 84
- design, 70
- dual monostable, 75–76, 77
- function, 69–70
- game sound, 85
- hysteresis, 83
- loudspeaker connection, 84
- monostable mode, 69, 71, 77, 79
- output time control, 80
- potential problems, 85–87
- TTL version, 70
- use of, 79–85
- values, 76–79
- variants, 70–76
- toggle, 43, 113
- total harmonic distortion (THD), 266
- transducer, 249–255, 257, 267
 - PT-2040PQ, 253
 - AC supply, 254
 - current, 254
 - design, 249
 - dynamic, 262
 - electromagnetic, 250, 253
 - formats, 251
 - frequency range, 251
 - function, 249
 - limitations, 253–254
 - measurement location, 253
 - piezoelectric, 249, 250, 253
 - potential problems, 254

- self-drive transducer circuit, 254
- sound intensity, 254
- sound pressure, 251–252
- surface-mount, 251
- ultrasonic, 250
- unweighted values, 253
- use of, 254
- values, 251–254
- variants, 250–251
- voltage, 254
- volume control, 254
- weighted sounds values, 252–253
- transformer, isolation, 34
- transistor, bipolar, 1
- transistor-transistor logic, 93
- transistor-transistor logic protocol, 70
- transparent, 108, 114
- transreflective LCD, 161
- tri-state encoder, 142
- triac, 1, 7, 11, 15–24, 25, 34
 - BTA208X-1000B, 16
 - BTB04-600SL, 16
 - MAC97A6, 16
 - breakover voltage, 20
 - charge storage, 23
 - design, 17–23
 - function, 15–16
 - other drivers, 22
 - potential problems, 24
 - quadrants, 17–18
 - snubberless, 23
 - switching AC, 20–22
 - symbol variants, 15–16
 - testing, 19–20
 - triggered by diac, 22
 - values, 23–24
 - variants, 23
- triac-based dimmer, 179
- trigger pin, 69
- trimmer, 40, 254
- truth table, 90
- TTL (see transistor-transistor logic)
- tweeter, 269
- twisted nematic (TN) LCD, 161
- two-wire chips, 64

U

- ultrasonic transducer, 250
- ultraviolet output, 222
- unity gain frequency, 53
- universal shift register, 125
- unweighted sound values, 253
- up/down protocol, 62, 63
- Upper State Transition Voltage (USTV), 43

V

- Vactrol, 35
- vacuum-fluorescent display, (VFD), 239–242
 - anode in, 240
 - cathode, 240
 - character sets and pictorial design, 242
 - color, 241
 - comparisons, 242
 - design, 239–240
 - function, 239
 - modern application, 241
 - potential problems, 242
 - use of, 240–241
 - variants, 241–242
- variable frequency drive, 239
- velocity, 262
- vent, 270
- VFD (see vacuum-fluorescent display)
- view angle, 208
- volatile memory, 61
- voltage, 259
 - breakdown, 1, 4, 20, 184
 - breakover, 4, 15, 20
 - dual, 39
 - falling, 43
 - forward, 212
 - ignition, 184
 - input, 155
 - input offset, 43, 53
 - maintaining, 185
 - in multiplexers, 155
 - operating, 259

- rated, 259
- reference, 70
- rising, 43
- starting/striking, 184
- threshold, 207
- transducer, 254
- voltage amplification, 53
- voltage amplifier, 49
- voltage converter, 39
- voltage divider, 71
- voltage divider mode, 65
- voltage gain, 43
- voltage level, 69
- voltage regulator, 59
- voltage spike, 72
- voltage transition, 69

W

- wafers, 219
- water clear, 205
- wavelength, 262
- wavelength coherence, 200
- wedge base, 177
- weighted output, 134
- weighted sound values, 252–253
- window comparator, 42, 46
- wiper, 60
- wiper resistance, 65
- wireless speaker, 271
- woofer, 269

X

- XNOR gate, 90, 91, 98, 100, 102, 103
- XOR gate, 90, 91, 92, 97, 98, 100, 102, 103

Z

- zero point finder, 47
- zero value, 49
- zero-crossing relay, 25, 27
- zero-crossing signal, 66

About the Authors

Charles Platt is a contributing editor and regular columnist for *Make:* magazine, where he writes about electronics. He is the author of the highly successful introductory hands-on book, *Make: Electronics*, and its sequel, *Make: More Electronics*. His science fiction novels are currently being reissued by Stairway Press.

Platt was a Senior Writer for *Wired* magazine and has written various computer books. As a prototype designer, he created semi-automated rapid cooling devices with medical applications, and air-deployable equipment for first responders. He was the sole author of four mathematical-graphics software packages, and has been fascinated by electronics since he put together a telephone answering machine from a tape recorder and military-surplus relays at age 15. He lives in a Northern Arizona wilderness area, where he has his own workshop for prototype fabrication and projects that he writes about for *Make:* magazine.

Fredrik Jansson is a physicist from Finland, with a PhD from Åbo Akademi University. He is currently living in the Netherlands, where he works on swarm robotics and simulates sea animals in the computational science group at the University of Amsterdam. Fredrik has always loved scavenging discarded household electronics for parts, and is a somewhat inactive radio amateur with the call sign OH1HSN. He also fact-checked Charles Platt's previous book, *Make: More Electronics*.

Colophon

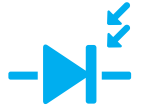
The cover and body font is Myriad Pro, the heading font is Benton Sans, and the code font is Ubuntu Mono.

Make:

VOL. 3

Charles Platt and Fredrik Jansson

Encyclopedia of Electronic Components



Sensors



**Location • Presence • Proximity • Orientation •
Oscillation • Force • Load • Human Input • Liquid and
Gas Properties • Light • Heat • Sound • Electricity**

Encyclopedia of Electronic Components

Sensors

Want to know how to use an electronic component? This third book of a three-volume set includes key information on electronic sensors for your projects—complete with photographs, schematics, and diagrams. You'll learn what each one does, how it works, why it's useful, and what variants exist. No matter how much you know about electronics, you'll find fascinating details you've never come across before.

Convenient, concise, well-organized, and precise

Perfect for teachers, hobbyists, engineers, and students of all ages, this reference puts reliable, fact-checked information right at your fingertips. Beginners will quickly grasp important concepts, and more experienced users will find the specific details their projects require.

- **Unique:** the first and only multivolume encyclopedia of electronic components, distilled into three separate volumes
- **Instructive:** each component description provides details about substitutions, common problems, and workarounds
- **Incredibly detailed:** includes information distilled from hundreds of sources
- **Easy to browse:** parts are clearly organized by component type
- **Authoritative:** fact-checked by experts to ensure current and accurate information
- **Reliable:** a more consistent source of information than online sources, product datasheets, and manufacturer's tutorials
- **Comprehensive:** Volume 1 covers power, electromagnetism, and discrete semiconductors; Volume 2 includes integrated circuits, and light and sound sources; Volume 3 covers all major types of sensing components

Charles Platt is the author of *Make: Electronics* and *Make: More Electronics*. He is a former senior writer for *Wired* magazine, and is a contributing editor to *Make: magazine*, for which he writes a column on electronics.

Fredrik Jansson is a physicist from Finland, with a PhD from Åbo Akademi University. He is currently living in The Netherlands, where he works on swarm robotics and simulates sea animals in the Computational Science group at the University of Amsterdam. Fredrik has always loved scavenging discarded household electronics for parts, and is a somewhat inactive radio amateur with the call sign OH1HSN.

US \$29.99 CAN \$34.99
ISBN: 978-1-4493-3431-4



5 2 9 9 9

Make:
makezine.com

Encyclopedia of Electronic Components Volume 3

Charles Platt and Fredrik Jansson



Encyclopedia of Electronic Components, Volume 3

by Charles Platt

Copyright © 2016 Charles Platt. All rights reserved.

Printed in Canada.

Published by Maker Media, Inc., 1160 Battery Street East, Suite 125, San Francisco, CA 94111.

Maker Media books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (<http://safaribooksonline.com>). For more information, contact O'Reilly Media's institutional sales department: 800-998-9938 or corporate@oreilly.com.

Editor: Brian Jepson

Production Editor: Melanie Yarbrough

Copyeditor: Christina Edwards

Proofreader: Charles Roumeliotis

Indexer: Charles Platt

Interior Designer: David Futato

Cover Designer: Karen Montgomery

Illustrator: Charles Platt

April 2016: First Edition

Revision History for the First Edition

2016-04-05: First Release

See <http://oreilly.com/catalog/errata.csp?isbn=9781449334314> for release details.

Make:, Maker Shed, and Maker Faire are registered trademarks of Maker Media, Inc. The Maker Media logo is a trademark of Maker Media, Inc. *Encyclopedia of Electronic Components, Volume 3* and related trade dress are trademarks of Maker Media, Inc.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and Maker Media, Inc. was aware of a trademark claim, the designations have been printed in caps or initial caps.

While every precaution has been taken in the preparation of this book, the publisher and authors assume no responsibility for errors or omissions, or for damages resulting from the use of the information contained herein.

978-1-449-33431-4

[TI]

To Brian Jepson

Table of Contents

Preface xix

1. GPS 1

 What It Does 1

 Schematic Symbol 1

 GPS Segments 1

 How It Works 2

 Variants 2

 Values 3

 How to Use It 3

 Pulse per Second Output 4

 What Can Go Wrong 4

 Electrostatic Discharge 4

 Failure to Ground Properly 4

 Cold Joints 4

 Restricted Availability 4

 Inability to Detect Satellites 4

 Exceeding Maximum Velocity or Altitude 4

2. magnetometer 5

 What It Does 5

 Schematic Symbol 5

 IMU 5

 Applications 6

 How It Works 6

 Magnetic Fields 6

 Earth's Axes 7

 Coil Magnetometer 8

 Hall Effect and Magnetoresistance 8

 Variants 9

How to Use It	9
What Can Go Wrong	9
Bias	9
Mounting Errors	10
3. object presence sensor	11
What It Does	11
Schematic Symbol	12
Variants	12
Optical Detection	12
Transmissive Optical Sensors	13
Retroreflective Optical Sensors	15
Magnetic Sensors	16
Reed Switch	17
Reed Switch Variants	17
Reed Switch Values	18
How to Use a Reed Switch	18
Hall-Effect Sensor	18
How a Hall-Effect Sensor Works	19
Hall-Effect Sensor Variants	19
Other Applications	20
Values	20
How to Use a Hall-Effect Sensor	20
Configuration of Object Presence Sensors	21
Linear Motion	21
Sensing by Interruption	21
Angular Motion	21
Sensor Comparisons	22
Advantages of Optical Presence Sensors	22
Disadvantages of Optical Presence Sensors	22
Advantages of a Reed Switch	22
Disadvantages of a Reed Switch	22
Advantages of a Hall Effect Sensor	22
What Can Go Wrong	23
Optical Sensor Issues	23
Reed Switch Issues	23
4. passive infrared sensor	25
What It Does	25
Schematic Symbols	25
Applications	25
How It Works	26
Pyroelectric Detector	26
Elements	26
Lenses	27
Variants	29

What Can Go Wrong	30
Temperature Sensitivity	30
Detector Window Vulnerability	30
Moisture Vulnerability	30
5. proximity sensor	31
What It Does	31
Schematic Symbols	31
Applications	32
Variants	32
Ultrasound	32
Infrared	32
Relative Advantages	33
Ultrasonic Examples	33
Imports	33
Individual Elements	34
Infrared Examples	34
Trends in Infrared Proximity Sensing	35
Capacitive Displacement Sensor	36
Applications	37
How It Works	37
Sources of Error	37
Values	38
What Can Go Wrong with Optical and Ultrasound Proximity Sensors	38
Object Too Close	38
Multiple Signals	38
Inappropriate Surfaces	38
Environmental Factors	38
Deterioration of LEDs	38
6. linear position sensor	39
What It Does	39
Applications	39
Schematic Symbol	39
How It Works	39
Linear Potentiometer	40
Magnetic Linear Encoders	41
Optical Linear Encoders	41
Linear Encoder Applications	42
Linear Variable Differential Transformers	42
What Can Go Wrong	43
Mechanical Issues	43
LED Longevity	43
7. rotary position sensor	45
What It Does	45

Applications	45
Schematic Symbol	46
Potentiometers	46
Arc-Segment Rotary Potentiometer	46
End Stops	46
Multiturn Rotary Potentiometer	46
Magnetic Rotary Position Sensor	47
Rotary Position Sensing Chips	48
Rotary Encoders	48
Optical Rotary Encoders	48
Optical Products	49
Computer Mouse Principles	50
Rotational Speed	50
Absolute Position	51
The Gray Code	51
Magnetic Rotary Encoders	52
How to Use It	53
What Can Go Wrong	53
Wiring Errors	53
Coding Errors	53
Ambiguous Terminology	54
8. tilt sensor	55
What It Does	55
Schematic Symbol	56
How It Works	56
Simplified Version	57
Applications	57
Variants	58
Mercury Switches	58
Pendulum Switch	58
Magnetization	59
Tilt Sensors	59
Two-Axis Tilt Sensors	59
Values	60
How to Use It	61
What Can Go Wrong	61
Contact Erosion	61
Random Signals	61
Environmental Hazard	61
Requirement for Gravity	61
Requirement for Stability	62
9. gyroscope	63
What It Does	63
Schematic Symbol	63

IMU	63
Applications	63
How It Works	64
Vibrating Gyroscope	64
Variants	66
IMUs	66
Values	67
How to Use It	67
What Can Go Wrong	67
Temperature Drift	67
Mechanical Stress	68
Vibration	68
Placement	68
10. accelerometer	69
What It Does	69
IMU	69
Schematic Symbols	70
Applications	70
How It Works	70
Gravity and Free Fall	71
Rotation	71
Calculation	72
Variants	72
Values	73
What Can Go Wrong	74
Mechanical Stress	74
Other Problems	74
11. vibration sensor	75
What It Does	75
Schematic Symbols	75
Variants	75
Pin-and-Spring	76
Piezoelectric Strip	76
Chip-Based Piezoelectric	77
“Mousetrap” Type	77
Magnetic	77
Mercury	78
Values	78
Primary Variables	78
Dynamic Attributes	79
How to Use It	79
What Can Go Wrong	79
Long Cable Runs	79
Interference	79

Correct Grounding	80
Fatigue Failure	80
12. force sensor	81
What It Does	81
Applications	81
Schematic Symbol	82
How It Works	82
Strain Gauge	82
Wheatstone Bridge Circuits	83
Wheatstone Bridge Errors	84
Strain-Gauge Amplification	84
Other Strain-Gauge Modules	85
Plastic-Film Force Sensors	85
How to Use It	86
Values	87
Film-Based Force Sensors for User Input	87
Specifications for Film-Based Force Sensors	87
Strain Gauges	88
What Can Go Wrong	88
Soldering Damage	88
Bad Load Distribution	88
Water Damage	88
Temperature Sensitivity	88
Leads Too Long	88
13. single touch sensor	89
What It Does	89
Applications	90
Schematic Symbols	90
How It Works	90
How to Use It	91
Obtaining Touch Pads	91
Individual Touch Pad	91
Wheels and Strips	92
Design Considerations	92
What Can Go Wrong	93
Insensitive to Gloves	93
Stylus Issues	93
Conductive Ink	93
14. touch screen	95
What It Does	95
Schematic Symbol	95
Variants	95
Resistive Sensing	95

Capacitive Sensing	96
Screens Available as Components	97
15. liquid level sensor	99
What It Does	99
Schematic Symbols	99
Applications	99
How It Works	100
Binary-Output Float Sensor	100
Analog-Output Float Sensor	101
Incremental-Output Float Sensor	101
Displacement Level Sensors	102
Ultrasonic Level Sensors	102
Reservoir Weight	103
Pressure Sensing	103
What Can Go Wrong	104
Turbulence	104
Tilting	104
16. liquid flow rate sensor	105
What It Does	105
Schematic Symbols	105
Paddlewheel Liquid Flow Rate Sensors	105
Turbine Flow Rate Sensors	106
Limitations of Paddlewheels and Turbines	107
Thermal Mass Liquid Flow Rate Sensor	107
Sliding Sleeve Liquid Flow Switch	108
Sliding Plunger Liquid Flow Switch	108
Ultrasonic Liquid Flow Rate Sensor	108
Magnetic Liquid Flow Sensor	108
Differential Pressure Liquid Flow Meter	109
What Can Go Wrong	109
Vulnerability to Dirt and Corrosive Materials	109
17. gas/liquid pressure sensor	111
What It Does	111
Schematic Symbols	111
Applications	111
Design Considerations	112
Units	112
How It Works	112
Basic Sensing Elements	112
Relative Measurement	113
Variants	114
Ambient Air Pressure	114
Altitude	114

Gas Pressure	115
What Can Go Wrong	116
Vulnerability to Dirt, Moisture, and Corrosive Materials	116
Light Sensitivity	116
18. gas concentration sensor	117
What It Does	117
Schematic Symbol	117
Semiconductor Gas Sensors	117
Oxygen Sensors	119
Humidity Sensors	119
Dew-Point Sensor	120
Absolute Humidity Sensors	120
Relative Humidity Sensors	120
Humidity Sensor Output	121
Analog Humidity Sensor	121
Design Considerations	122
Digital Humidity Sensor	122
What Can Go Wrong	122
Contamination	122
Recalibration	123
Soldering	123
19. gas flow rate sensor	125
What It Does	125
Applications	125
Schematic Symbol	125
How It Works	126
Anemometer	126
Mass Flow Rate Sensing	127
Applications	128
Units	128
Measuring Higher Volumes	128
Output	128
What Can Go Wrong	129
20. photoresistor	131
What It Does	131
Schematic Symbol	131
How It Works	132
Construction	132
Variants	132
Photoresistors in Optical Isolators	132
Values	133
Comparisons with a Phototransistor	133
How to Use It	133

Choosing a Series Resistor	134
What Can Go Wrong	134
Overload	134
Excessive Voltage	134
Confusion Among Components	134
21. photodiode	135
What It Does	135
Schematic Symbols	135
Applications	135
How It Works	135
Variants	136
PIN Photodiodes	136
Avalanche Diodes	136
Packages	136
Wavelength Range	137
Photodiode Arrays	137
Output Options	137
Specific Variants	137
Values	138
How to Use It	139
What Can Go Wrong	140
22. phototransistor	141
What It Does	141
Schematic Symbols	141
Applications	142
How It Works	142
Variants	142
Optional Base Connection	142
Photodarlington	142
PhotoFET	142
Values	143
Behavior Compared to Other Light Sensors	143
Binning	144
How to Use It	144
Output Calculation	145
What Can Go Wrong	145
Visual Classification Errors	145
Output Out of Range	145
23. NTC thermistor	147
What It Does	147
Schematic Symbols	147
Applications	148
How an NTC Thermistor Works	148

Output Conversion for Temperature Sensing	149
Choosing a Series Resistor	149
Wheatstone Bridge Circuit	150
Deriving the Temperature Value	150
Inrush Current Limiter	150
Restart	151
Thermistor Values	152
Time and Temperature	152
Resistance and Response	152
Kilohms and Kelvin	152
Reference Temperature	152
Reference Resistance	152
Dissipation Constant	152
Temperature Coefficient	152
Thermal Time Constant	152
Tolerance	153
Temperature Range	153
Switching Current	153
Power Limitations	153
Interchangeability	153
What Can Go Wrong	153
Self-Heating	153
Heat Dissipation	153
Lack of Heat	153
Addendum: Comparison of Temperature Sensors	153
NTC Thermistor	154
PTC Thermistor	154
Thermocouple	154
Resistance Temperature Detector	154
Semiconductor Temperature Sensor	155
24. PTC thermistor	157
What It Does	157
Schematic Symbols	157
PTC Overview	158
Silistor for Temperature Measurement	158
RTDs	159
Nonlinear PTC Thermistors	159
Over-Temperature Protection	159
Over-Current Protection	160
PTC Inrush Current Limiting	161
PTC Thermistor for Lighting Ballast	162
PTC Thermistor as a Heating Element	162
What Can Go Wrong	163
Self-Heating	163
Heating Other Components	163

25. thermocouple	165
What It Does	165
Schematic Symbol	166
Thermocouple Applications	166
How a Thermocouple Works	167
Thermocouple Details	168
How to Use It	168
Types of Thermocouples	169
Seebeck Coefficients	169
Chips for Output Conversion	170
Thermopile	171
What Can Go Wrong	171
Polarity	171
Electrical Interference	171
Metal Fatigue and Oxidation	171
Using the Wrong Type	171
Heat Damage from Creating a Thermocouple	172
26. RTD (resistance temperature detector)	173
What It Does	173
RTD Attributes	173
Schematic Symbol	174
Applications	174
How It Works	174
Variants	175
Wiring	175
RTD Probe	176
Signal Conditioning	176
What Can Go Wrong	176
Self-Heating	176
Insulation Affected by Heat	176
Incompatible Sensing Element	176
27. semiconductor temperature sensor	177
What It Does	177
Semiconductor Temperature Sensor Applications	178
Schematic Symbol	178
Attributes	178
How It Works	179
CMOS Sensors	179
Multiple Transistors	179
PTAT and the Brokaw Cell	180
Variants	180
Analog Voltage Output	181
Analog Current Output	182
Digital Output	183

CMOS Semiconductor Temperature Sensors	185
What Can Go Wrong	185
Different Temperature Scales	185
Interference in Cable Runs	185
Latency	186
Processing Time	186
28. infrared temperature sensor	187
What It Does	187
Applications	188
Schematic Symbol	188
How It Works	188
Thermopile	189
Temperature Measurement	190
Variants	190
Surface-Mount Specifications	191
Sensor Arrays	191
Values	191
Temperature Range	191
Field of View	192
What Can Go Wrong	192
Inappropriate Field of View	192
Reflective Objects	192
Glass Obstruction	192
Multiple Heat Sources	192
Thermal Gradients	192
29. microphone	193
What It Does	193
Schematic Symbol	193
How It Works	194
Carbon Microphone	194
Moving-Coil Microphone	194
Condenser Microphone	194
Electret Microphone	195
MEMS Microphone	195
Piezoelectric Microphone	196
Values	196
Sensitivity	196
Directionality	197
Frequency Response	197
Impedance	198
Total Harmonic Distortion	198
Signal-to-Noise Ratio	198
What Can Go Wrong	198
Cable Sensitivity	198

Noisy Power Supply	198
30. current sensor	199
What It Does	199
Applications	199
Ammeter	199
Schematic Symbol	200
Ammeter Wiring	200
Series Resistor	200
Current-Sense Resistors	201
Voltage Measurement	202
Hall-Effect Current Sensing	202
What Can Go Wrong	203
Confusing AC with DC	203
Magnetic Interference	203
Incorrect Meter Wiring	203
Current Out of Range	203
31. voltage sensor	205
What It Does	205
Applications	205
Volt Meter	205
Schematic Symbol	206
Volt Meter Wiring	206
How It Works	206
Load-Related Inaccuracy	207
Bar Graph	207
What Can Go Wrong	208
Confusing AC with DC	208
High Circuit Impedance	208
Voltage Out of Range	208
Voltage Relative to Ground	208
Appendix A. Sensor Output	209
Glossary	217
Index	221

Preface

This third and final volume of the *Encyclopedia of Electronic Components* is devoted entirely to sensors.

Two factors have caused very significant changes in the field of sensors since the 1980s. First, features such as antilock braking, airbags, and emissions controls stimulated the development of low-priced sensors for automotive applications. Many of these sensors were fabricated in silicon as MEMS (microelectromechanical) devices.

The second wave began in 2007 when MEMS sensors were installed in the iPhone. A modern phone may contain almost a dozen different types of sensors, and their size and price have been driven down to a point that would have been unimaginable 20 years previously.

Many MEMS sensors are now as cheap as basic semiconductor components such as a voltage regulator or a logic chip, and they are easy to use in conjunction with microcontrollers. In this Encyclopedia, we have allocated significant space to this segment of the market, hoping that the specific products that we have chosen will remain popular and available for at least the next decade.

In addition, we have devoted space to older components where durability has been proven.

Purpose

While much of the information in this volume can be found dispersed among datasheets, introductory texts, Internet sites, and technical briefings published by manufacturers, we believe there is a real need for a durable resource that assembles all the relevant data in one place, properly organized and verified, including details that may be hard to find elsewhere.

This volume may also serve a useful purpose by attempting to categorize and classify components in a field that is remarkably chaotic. For example, is an *object presence sensor* different from a *proximity sensor*? Some manufacturers seem to think so; others disagree. Understanding the distinctions and the underlying principles can be important if you are trying to decide which sensor to use.

Sensor terminology can also be confusing. To take another example, what is the difference between a *reflective interrupter*, a *reflective object sensor*, a *reflective optical sensor*, a *reflective photointerrupter*, and an *opt-pass sensor*? These terms are used in various datasheets to describe components that are all *retroreflective sensors*. Understanding the proliferating variety of terminology can be essential if you simply want to find something in a product index.

Organization

As in volumes 1 and 2, this volume is organized by subject. For example, if you want to measure temperature, you'll find the entries for a thermistor and a thermocouple next to each other, in an entire section devoted to the sensing of heat. This will help you to compare capabilities and choose the component that best suits your application.

The subject path leading to each sensor is shown at the top of the first page of each entry. For gas flow rate, for instance, you would follow this path:

fluid > gas > flow rate

Note that the word "fluid" is properly used to include gases as well as liquids.

Exceptions and Conflicts

Unfortunately, some sensors are not easily categorized. There are four problems in this area.

1. What Does a Sensor Really Sense?

A GPS chip is a radio receiver, picking up transmissions from satellites. Does this mean it should be categorized as a sensor of radio waves? No, its purpose is to tell you your location. Therefore, it is categorized as a location sensor. This leads to the first general rule: sensors are categorized by their primary purpose. Secondary purposes may be found in the index.

2. How Many Sensors Are in a Sensor?

Many surface-mount chips perform more than one sensing function. For example, an inertial measurement unit (often identified by its acronym, *IMU*) can contain three gyroscope sensors and three accelerometers—and may contain three magnetometers, too. How should it be categorized?

The answer is that an IMU will be mentioned in more than one entry in the Encyclopedia, because it performs more than one function; but it will not have its own separate entry,

because each entry in the Encyclopedia is for a single primary sensing function.

The names of multisensor chips are, of course, included in the index.

3. How Many Stimuli Can One Sensor Sense?

A single sensing element may be used in multiple different types of sensors. The most notable example is the Hall-effect sensor, which can be found in magnetometers, object presence sensors, speed sensors, current sensors, and dozens more. Modern automobiles can contain Hall-effect sensors everywhere from the ignition system to the trunk-locking mechanism. If you are using a hard drive with rotating platters, it probably contains a Hall-effect sensor to monitor the speed of rotation. If you have a generic computer keyboard, each keypress is probably detected with a Hall-effect sensor.

Bearing this in mind, how should a Hall-effect sensor be classified? And where should you expect to find an explanation of how it works?

The answer is that where different types of components contain the same type of sensing element, the entry for each component will include a cross-reference to one location where the sensing element is explained in detail.

This location will be chosen for its relevance. Thus, Hall-effect sensors are explained in the entry for **object presence** sensors, because this is their primary function. While it is true that a Hall-effect sensor works by detecting a magnetic field, that is not its most common application.

4. Too Many Sensors!

Wikipedia lists more than [100 general types of sensors](#), and even that list is probably not complete. Consequently, we had to pick and choose. Some of the decisions may seem arbitrary, but all of them were made on the grounds of practicality. There were three principles for deciding what to include and what to leave out.

1. Is it a component? We are more interested in board-mounted components than in packaged products that happen to contain sensors. For instance, a thermocouple is often enclosed in a tubular steel probe, and its wire is often plugged in to a specially designed meter that displays temperature. While we do include a photograph of a probe, we are primarily interested in the welded wires of the thermocouple inside it.
2. How much does it cost? An industrial ultrasonic sensor to check items on a factory conveyor belt will be sealed into a module with a waterproof grommet around a shielded cable—which is all very nice, but will not be very affordable. This Encyclopedia is more interested in board-mountable components for one-tenth of the price.
3. How many people are likely to want it? The stock of each type of sensor was checked on component vendor sites. If a sensor wasn't in the inventory, or if only a couple of variants were stocked, we concluded that the limited demand probably didn't justify including it here. For example, a Ferraris acceleration sensor responds to eddy currents in a rotating motor shaft, as a way of measuring vibration in the shaft. This is a really interesting device, but is unlikely to be on most people's shopping lists.

Volume Contents

Having explained the organization of this book and our decisions to include or omit various components, we now present a summary of the contents of all three Encyclopedia volumes:

Volume 1

Power; electromagnetic devices; discrete semi-conductors.

The *power* category includes sources of electricity and methods to distribute, store, interrupt, convert, and regulate power. The *electromagnetism* category includes devices that exert force linearly, and others that create a turning force. *Discrete semiconductors* include the primary types of diodes and transistors. See [Figure P-1](#) for a contents listing.

Primary Category	Secondary Category	Component Type
power	source	battery
		connection
		jumper
		fuse
		pushbutton
		switch
		rotary switch
		rotational encoder
	moderation	relay
		resistor
		potentiometer
		capacitor
		variable capacitor
	conversion	inductor
		AC-AC transformer
		AC-DC power supply
		DC-DC converter
		DC-AC inverter
	regulation	voltage regulator
electro-magnetism	linear output	electromagnet
		solenoid
	rotational output	DC motor
		AC motor
		servo motor
discrete semi-conductor	single junction	diode
		unijunction transistor
	multi-junction	bipolar transistor
		field-effect transistor

Figure P-1 The subject-oriented organization of categories and entries in Volume 1 of this Encyclopedia.

Volume 2

Thyristors (SCRs, diacs, and triacs); integrated circuits; light sources, indicators, and displays; and sound sources.

Integrated circuits are divided into analog and digital components. *Light sources, indicators, and displays* are divided into reflective displays, single sources of light, and displays that emit light. *Sound sources* are divided into those that create sound, and those that reproduce sound. A contents listing for Volume 2 appears in [Figure P-2](#).

Volume 3

All the most common types of sensing devices, including those that detect location, presence, proximity, orientation, oscillation, force, load, human input, liquid properties, gas types and concentrations, pressure, flow rate, light, heat, sound, and electricity. A contents listing for Volume 3 appears in [Figure P-3](#).

Method

Reference Versus Tutorial

As its title suggests, this is a reference book, not a tutorial. A tutorial such as *Make: Electronics* begins with elementary concepts and builds sequentially toward concepts that are more advanced. A reference book assumes that you may dip into the text at any point, learn what you need to know, and then put the book aside. If you choose to read it straight through from beginning to end, you will find some repetition, as each entry is intended to be self-sufficient, requiring minimal reference to other entries.

Theory and Practice

This book is oriented toward practicality rather than theory. We assume that the reader mostly wants to know how to use electronic components, rather than why they work the way they do. Consequently we do not include detailed proofs of formulae or definitions rooted in electrical theory.

Primary Category	Secondary Category	Component Type
discrete semi-conductor	thyristor	SCR
		diac
		triac
integrated circuit	analog	solid-state relay
		optocoupler
		comparator
		op-amp
		digital potentiometer
		timer
	digital	logic gate
		flip-flop
		shift register
		counter
		encoder
		decoder
		multiplexer
light source, indicator or display	reflective	LCD
	single source	incandescent lamp
		neon bulb
		fluorescent light
		laser
		LED indicator
		LED area lighting
	multi-source or panel	LED display
		vacuum-fluorescent
		electroluminescence
sound source	audio alert	transducer
		audio indicator
	reproducer	headphone
		speaker

Figure P-2 The subject-oriented organization of categories and entries in Volume 2.

Primary Category	Attribute to be Sensed	Type of Sensor
spatial	location	GPS
		magnetometer
	presence	object presence
		passive infrared
	distance	object proximity
		linear position
	orientation	rotary position
		tilt
		gyroscope
accelerometer		
mechanical	oscillation	vibration
	force	force
	human input	single touch
		touch screen
fluid	liquid	liquid level
		liquid flow rate
	gas/liquid	pressure
	gas	gas concentration
		gas flow rate
radiation	light	photoresistor
		photodiode
		phototransistor
	heat	NTC thermistor
		PTC thermistor
		thermocouple
		RTD
		semiconductor
		infrared temperature
		sound
electricity	metering	current
		voltage

Figure P-3 The subject-oriented organization of categories and entries in Volume 3.

Sensor Output

In Volumes 1 and 2 of the *Encyclopedia*, each entry included hints on how to use a component. However, many sensors have identical

forms of output, which are processed in a similar way. To avoid repetition, general guidance for using nine principal types of sensor outputs has been placed in [Appendix A](#) at the back of this volume.

For example, many sensors provide an analog voltage output that varies with the phenomenon that is being sensed. In [Appendix A](#), you will find suggestions on how to adjust the range of the output, if necessary, or how to digitize it with an analog-to-digital converter.

You will also find a comparison between serial protocols such as I2C and SPI, both of which are commonly used when a microcontroller communicates with a digital sensor via a bus.

Glossary

In the world of sensors, many terms tend to recur. *Hysteresis* is one; *MEMS* is another. Rather than define these terms repeatedly, some quick definitions are gathered in a Glossary. Please remember the existence of the glossary if you encounter a term that is unfamiliar. See [Glossary](#).

In many instances, terms that are italicized in the text are defined in the glossary.

Typographical Conventions

Within each entry, **bold type** is used for the first occurrence in each entry of the name of a component that has its own entry elsewhere. Other important electronics terms or component names may be presented in *italics*.

The names of components, and the categories to which they belong, are all set in lowercase type, except where a term is normally capitalized because it is an acronym or a trademark, or contains a proper noun. The term *Hall effect*, for instance, has an initial cap because it is named after a person named Hall. The term *GPS* is all in caps, because it is an acronym; but *psi* (meaning pounds per square inch) remains in lowercase, because even though it is an acronym, the lowercase form is more common.

The situation is different when specifying units that are named after electrical pioneers. All of these units should be lowercased when spelled out. Thus, when referring to the SI unit of force, it is “the newton.” However, where a unit named after a person is abbreviated, the abbreviation is capitalized, as in N for newtons, Hz for hertz, Pa for pascals, and A for amperes.

Mathematical Syntax

In mathematical formulae, we have used the style that is common in programming languages. The * (asterisk) is used as a multiplication symbol, while the / (forward slash) is used as a division symbol. Where some terms are in parentheses, they must be dealt with first. Where parentheses are inside parentheses, the innermost ones must be dealt with first. Consider this example:

$$A = 30 / (7 + (4 * 2))$$

You would begin by multiplying 4 times 2, to get 8; then add 7, to get 15; then divide that into 30, to get the value for A, which is 2.

Visual Conventions

Figure P-4 shows the conventions that are used in the schematics in this book. A black dot always indicates a connection, except that to minimize ambiguity, the configuration at top-right is avoided, and the configuration at top-center is used instead. Conductors that cross each other without a black dot do not make a connection. The styles at bottom right are sometimes seen elsewhere, but are not used here.

All the schematics are formatted with pale blue backgrounds. This enables components such as switches, transistors, and LEDs to be highlighted in white, drawing attention to them and clarifying the boundary of the component. The white areas have no other meaning.

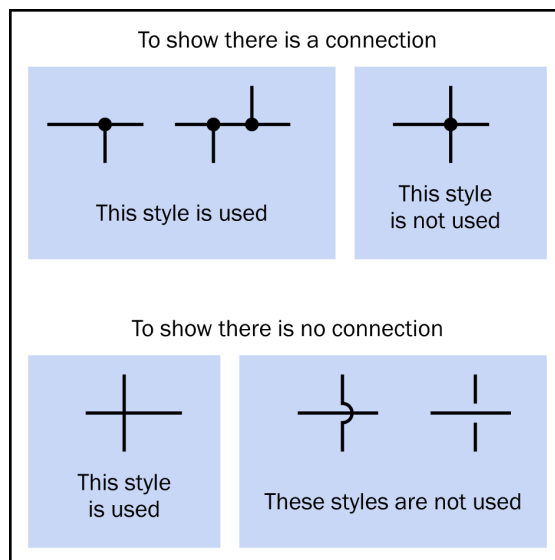


Figure P-4 Visual conventions that are used in the schematics in this book.

Units and Backgrounds

So long as the United States clings stubbornly to the habit of expressing dimensions in inches, there is a good argument to follow this custom in books intended for an American audience. With this in mind, Volumes 1 and 2 mostly avoided metric units of length. However, as time passed, the books were translated for use in many parts of the world where the inch is regarded as an anachronism.

Recognizing that we now have an international audience, we have used the metric system throughout this volume (with very few exceptions, such as a photograph of an American plumbing fixture that is designed to fit 3/4" pipe). For readers who are metrically impaired, here are some units of length, and their abbreviations:

- 1 nanometer (nm)
- 1 micrometer (μm) = 1,000nm
- 1 millimeter (mm) = 1,000μm
- 1 centimeter (cm) = 10mm
- 1 meter (m) = 100cm = 1,000mm

A micrometer is also known as a *micron*.

The basic conversion factor from meters to inches is 0.0254. Thus:

- 1 inch = 2.54cm = 25.4mm
- 1/1000 inch = 25.4μm

Sometimes 1/1000 inch is called a *mil*.

In many of the component photographs, a graph-paper background is included. Each square in these backgrounds is 1mm.

To avoid confusion, please remember that a few of these same component photographs appeared in books such as *Make: More Electronics*, where the background grid was in tenths of an inch. Captions to photographs in this volume will remind you that millimeters are now used.

Background colors in the photographs were chosen for contrast with the colors of the components, or for visual variety. They have no other significance.

Component Availability

The world of sensors is changing rapidly, and we have no way of knowing if a component will enjoy a long production run. We recommend checking availability at the following suppliers, which we used frequently during the preparation of the book:

- <http://www.mouser.com>
- <http://www.jameco.com>
- <http://www.sparkfun.com>
- <http://www.adafruit.com>

For obsolete parts, or those that are nearing the end of their commercial life, eBay can be very useful. Alternatively, new substitutions for old parts are often listed at <http://www.mouser.com>.

Issues and Errata

There are three situations where the reader and the writer may want to communicate with each other.

- We may want to tell you if the book contains a mistake of some significance. This is *us-informing-you* feedback.
- You may want to tell us if you think you found an error in the book. This is *you-informing-us* feedback.
- You may be having trouble making something work, and you don't know whether we made a mistake or you made a mistake. You would like some help. This is *you-asking-us* feedback.

Here's how you can deal with each of these situations.

Us Informing You

If you already registered your contact information in connection with *Make: Electronics* (second edition) or *Make: More Electronics*, you don't need to register again for updates relating to the Encyclopedia. If you have not already registered, here's how it works.

The only way you can be notified if there's an error in the book is if you supply your contact information. If we have your email address:

- You will be notified of any significant errors that are found in this book, and you will receive a correction.
- You will be notified if there is a completely new edition of this book, or of *Make: Electronics*, or any other books by Charles Platt. These notifications will be very rare.

Your contact information will not be used for any other purpose.

Simply send a blank email (or include some comments in it, if you like) to:

make.electronics@gmail.com

Please put REGISTER in the subject line.

You Informing Us

If you only want to report an error that you have found, it's really better to use the "errata" system maintained by our publisher. The publisher uses the "errata" information to fix the error in updates of the book.

If you feel sure that you found an error, please visit:

http://bit.ly/encyclopedia_electronic_components_v3

The web page will tell you how to submit errata.

You Asking Us

Our time is obviously limited, but if you have a question, a quick answer may be available. You can send email to make.electronics@gmail.com for this purpose. Please put the word HELP in the subject line.

Going Public

There are dozens of forums online where you can discuss this book and mention any problems you are having, but please be aware of the power that you have as a reader, and use it fairly. A single negative review can create a bigger effect than you may realize. It can certainly outweigh half-a-dozen positive reviews.

Responses in the past have been generally positive, but in a couple of cases people have been annoyed over small issues such as being unable to find a part online. Help is available on this kind of topic, if you need it. All you have to do is send a request to make.electronics@gmail.com.

Safari® Books Online



Safari Books Online is an on-demand digital library that delivers expert content in both book and video form from the world's leading authors in technology and business.

Technology professionals, software developers, web designers, and business and creative professionals use Safari Books Online as their primary resource for research, problem solving, learning, and certification training.

Safari Books Online offers a range of [plans and pricing](#) for [enterprise](#), [government](#), [education](#), and individuals.

Members have access to thousands of books, training videos, and prepublication manuscripts in one fully searchable database from publishers like O'Reilly Media, Prentice Hall Professional, Addison-Wesley Professional, Microsoft Press, Sams, Que, Peachpit Press, Focal Press, Cisco Press, John Wiley & Sons, Syngress, Morgan Kaufmann, IBM Redbooks, Packt, Adobe Press, FT Press, Apress, Manning, New Riders, McGraw-Hill, Jones & Bartlett, Course Technology, and hundreds [more](#). For more information about Safari Books Online, please visit us [online](#).

You can access the errata page at <http://bit.ly/encyclopedia-electronic-components-v3>.

Make: unites, inspires, informs, and entertains a growing community of resourceful people who undertake amazing projects in their backyards, basements, and garages. Make: celebrates your right to tweak, hack, and bend any technology to your will. The Make: audience continues to be a growing culture and community that believes in bettering ourselves, our environment, our educational system—our entire

world. This is much more than an audience, it's a worldwide movement that Make is leading. We call it the Maker Movement.

For more information about Make, visit us online:

Make: magazine: <http://makezine.com/magazine>

Maker Faire: <http://makerfaire.com>

Makezine.com: <http://makezine.com>

Maker Shed: <http://makershed.com>

To comment or ask technical questions about this book, send email to:

bookquestions@oreilly.com.

Acknowledgments

Datasheets and tutorials maintained by component manufacturers were considered the most trustworthy sources of information online. In addition, component retailers, college texts, crowd-sourced reference works, and hobbyist sites were used. The following books provided useful information:

Boylestad, Robert L. and Nashelsky, Louis: *Electronic Devices and Circuit Theory*, 9th edition. Pearson Education, 2006.

Braga, Newton C.: *CMOS Sourcebook*. Sams Technical Publishing, 2001.

Hoenig, Stuart A.: *How to Build and Use Electronic Devices Without Frustration, Panic, Mountains of Money, or an Engineering Degree*, 2nd edition. Little, Brown, 1980.

Horn, Delton T.: *Electronic Components*. Tab Books, 1992.

Horn, Delton T.: *Electronics Theory*, 4th edition. Tab Books, 1994.

Horowitz, Paul and Hill, Winfield: *The Art of Electronics*, 2nd edition. Cambridge University Press, 1989.

Ibrahim, Dogan: *Using LEDs, LCDs, and GLCDs in Microcontroller Projects*. John Wiley & Sons, 2012.

Kumar, A. Anand: *Fundamentals of Digital Circuits*, 2nd edition. PHI Learning, 2009.

Lancaster, Don: *TTL Cookbook*. Howard W. Sams & Co, 1974.

Lenk, Ron and Lenk, Carol: *Practical Lighting Design with LEDs*. John Wiley & Sons, 2011.

Lowe, Doug: *Electronics All-in-One for Dummies*. John Wiley & Sons, 2012.

Mims III, Forrest M.: *Getting Started in Electronics*. Master Publishing, 2000.

Mims III, Forrest M.: *Electronic Sensor Circuits & Projects*. Master Publishing, 2007.

Mims III, Forrest M.: *Timer, Op Amp, & Optoelectronic Circuits and Projects*. Master Publishing, 2007.

Predko, Mike: *123 Robotics Experiments for the Evil Genius*. McGraw-Hill, 2004.

Scherz, Paul: *Practical Electronics for Inventors*, 2nd edition. McGraw-Hill, 2007.

Williams, Tim: *The Circuit Designer's Companion*, 2nd edition. Newnes, 2005.

In addition, three individuals provided special assistance. Our editor, Brian Jepson, was immensely helpful in the development of this book. Philipp Marek reviewed the text for errors, and Erico Narita collaborated on the Photoshop work.

GPS

1

The acronym **GPS** properly refers to the entire Global Positioning System, including satellites and ground-based control installations. However, a *GPS sensor* consists of a surface-mount chip that processes signals from GPS satellites using a small rectangular antenna, often mounted on top of a *GPS chip*.

A *GPS module* is usually a small board on which a GPS sensor is mounted with additional components. A *GPS receiver* is a device including a data display and other features, such as memory, in addition to a GPS module. In casual colloquial usage, someone who refers to “a GPS” usually means a GPS receiver.

GPS is almost always capitalized without periods.

OTHER RELATED COMPONENTS

- **magnetometer** (see [Chapter 2](#))

What It Does

The Global Positioning System is a navigational aid jointly funded by the U.S. Department of Defense and the U.S. Department of Transportation, while being maintained by the U.S. Air Force. Signals from GPS satellites can be received and processed by modules in a wide variety of equipment ranging from aircraft to wristwatches. The signals provide location data, and may also be used as an accurate time reference.

Schematic Symbol

There is no specific schematic symbol for a GPS chip. It is likely to be shown as a box containing abbreviations that define pin functions, similar to any integrated circuit chip.

GPS Segments

The Global Positioning System consists of three segments:

The space segment

This originally required 24 communications satellites, but was revised in 2011 to require 27, to provide better global coverage. As of August 2015, there were actually 31 satellites in service, with additional “spares” that can be activated if necessary. The satellites occupy orbits 12,500 miles above the Earth, allowing each of them to circle the planet twice in 24 hours. Specifications are maintained [online](#).

The control segment

This includes a master ground-based control station, an alternate master control station, 12 command and control antennas, and 16 monitoring sites, all maintained by the U.S. Air Force.

The user segment

This consists of receiving devices, including those that are government-owned and those that are privately owned.

How It Works

Each satellite carries multiple atomic clocks that maintain precise time, and a pseudo-random number generator in the form of a linear-feedback **shift register** (see Volume 2).

A GPS receiver can distinguish the signals from at least four satellites by comparing their received pseudo-random bit sequences, and can compute the receiver's distance to each satellite by comparing the arrival times of satellite signals.

When a satellite appears above the horizon, it approaches a receiver. After passing overhead, it moves away. This relative motion causes a *Doppler shift* in the received frequency, which the receiver circuit must take into account.

GPS satellites transmit on several frequencies simultaneously. The one for civilian use is 1575.42 MHz, called L1. Another one, 1227.6 MHz, called L2, is reserved for military use.

Variants

A GPS chip generally processes input from an antenna and provides output through solder pads. The antenna is often integrated as a ceramic square or rectangle mounted above the chip, but many chips can also process input from an external antenna. [Figure 1-1](#) shows a GPS chip with a metal shield that is easily mistaken for an antenna. In [Figure 1-2](#), the GPS sensor does incorporate a ceramic antenna.

Some GPS chips contain flash memory for internal data logging, although this is not a standard feature.

Suppliers such as Adafruit and Sparkfun offer GPS modules mounted on breakout boards for

easier connection with other components, as shown in [Figure 1-2](#). Some breakout boards also include provision for battery backup with a button cell.



Figure 1-1 A GPS sensor. This surface-mount chip is hidden beneath a metal shield.

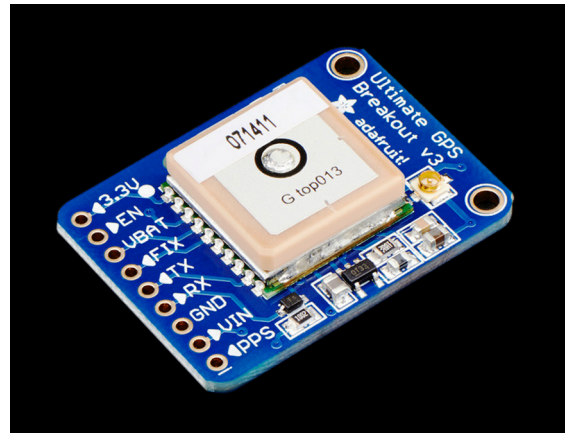


Figure 1-2 A breakout board offered by Adafruit, incorporating a GPS sensor.

GPS capability is almost always included in modern smartphones and tablets. It is used in handheld devices for navigation on foot, and in devices designed to be mounted in motor vehicles. Cars offer GPS capability as an option accessed via a built-in screen.

A *GPS tracker* is a device that may lack a display and simply logs its position to internal memory, from which data can be downloaded to a com-

puter later. Many (older) handheld GPS receivers have a connector giving access to a serial or USB port, and provide data in the same NMEA format as the GPS modules described below.

After the Global Positioning System became widely used, other competing systems were introduced. These include the European Galileo, the Russian GLONASS (an acronym for Global Navigation Satellite System), and the Chinese Beidou. As of 2015, GLONASS had become fully operational. Some receivers, including those in cellular phones, compare signals from GPS and GLONASS satellites to achieve higher accuracy.

Values

Sensitivity is expressed usually in dBm, meaning the power ratio in decibels (dB) of the measured power referenced to one milliwatt (mW).

Time to First Fix (TTFF) is the time required to obtain an initial fix from a satellite.

Number of channels is the number of satellites that a GPS receiver can track simultaneously. Early GPS receivers could sense only four channels. Modern units may be able to deal with 22.

Power consumption may be measured in milliwatts. For example, the FGPMOPA6H GPS standalone module by G.top claims power consumption of 82mW during acquisition of satellite signals and 66mW while tracking subsequently. At a typical voltage of 4VDC, the chipset consumes about 20mA and approximately 17mA, respectively.

Form factor. This is the size of a chip, often determined by the dimensions of the ceramic antenna on top of it. Dimensions may be 15mm x 15mm or larger.

Update rate. The number of position measurements per second. While 1 update per second is often sufficient, some chips generate updates more rapidly. The frequency of updates is expressed in hertz.

Output type. This is often TTL-level serial providing NMEA data. The baud rate can vary and is often selectable.

Supply voltage. Often below 5VDC.

Current consumption. Higher during satellite acquisition.

How to Use It

A GPS module requires only a DC power supply, and will start outputting data as soon as it has identified satellites that are currently within range.

Data provided by a GPS module uses a rather slow and primitive plain-ASCII protocol known as NMEA, developed by the National Marine Electronics Association. Each block of data is known as a **sentence**, and can be parsed independently of previous and subsequent sentences. The default transmission rate is 4,800 bits per second, using 8 bits to identify an ASCII character, no parity, and 1 stop bit. However, some GPS modules use a serial rate of 9,600 bps or faster.

A sentence begins with a two-letter abbreviation defining the type of device employing the sentence. For a GPS device, the abbreviation is GP. The sentence continues with another abbreviation of three characters or more, describing the type of data being transmitted, so that numeric values in the sentence can be interpreted correctly.

The remainder of a sentence consists of letters and numerals in plain ASCII, with values separated by commas. A sentence cannot contain more than 80 characters. A sentence will specify the latitude, longitude, and altitude of the GPS, and a value defining the time when the readings were derived from satellite signals. Some sentence data structures are proprietary, established by the device manufacturer, and will begin with the letter P.

A GPS device may send a variety of different sentence types in succession, to overcome the 80-character sentence length limit. Each sentence will be preceded with an identifier. Sentence types and data contents will be defined in a manufacturer's datasheet.

Output from GPS chips may be compatible with a microcontroller. Output from a GPS breakout board will almost certainly be compatible, and the board is likely to have its own voltage regulator. The microcontroller can receive serial data from the GPS chip, can stop the GPS chip via its enable pin, and can start and stop internal logging of data in flash memory on the GPS chip if this feature is available.

Code libraries are available online for a microcontroller, to enable it to receive and interpret serial data from a GPS device.

Pulse per Second Output

As the GPS positioning depends on calculating distances from the time it takes for a radio signal to travel, accurate timekeeping is needed. When a GPS receiver obtains a fix for its position, it will get a value for the current time as well. This makes GPS receivers usable to supply time and frequency standards. Most receiver modules report the time together with the position. Many also provide a special PPS output, which produces one pulse per second, precisely synchronized with the satellites in view.

The accurate time provided by the GPS receiver can be used to *discipline* a crystal oscillator. This means measuring the crystal oscillator frequency relative to a GPS-provided reference, and continuously adjusting the crystal frequency to keep it stable.

What Can Go Wrong

Problems generally affect chips and modules rather than devices.

Electrostatic Discharge

The patch antenna on a GPS chip connects with the chip via an RF input. If the antenna is subjected to an electrostatic discharge, the chip can be permanently damaged. Likewise, damage will result if a discharge is applied to the RF input, for example through a soldering iron. The chip must be grounded before any work is done involving the RF input.

Failure to Ground Properly

The ground solder pad on a chip, or the ground pin on a breakout board, must make contact before voltage is applied to other pads or pins.

Cold Joints

The patch antenna on a GPS chip can function as a heat sink, increasing the risk of cold solder joints when the chip is being mounted on a board.

Restricted Availability

U.S. regulations limit the export of some GPS devices capable of rapid positional updates that could be usable in military aircraft or missiles. Other restrictions also apply. Suppliers may be inconsistent in their refusal to allow purchase of restricted items outside of the United States.

Inability to Detect Satellites

Any GPS device may fail to detect satellite signals if its view of the sky is obstructed. Reception is usually possible through window glass, but may not be available through walls, a roof, heavy tree cover, and other natural obstructions.

Exceeding Maximum Velocity or Altitude

Security regulations limit the capability of GPS devices to function above 60,000 feet or at greater than 1,200 miles per hour. Outside of these limits, the GPS device will not provide data. This may affect applications in amateur rocketry or high-altitude balloons.

magnetometer



This entry deals only with magnetic sensors that respond to the Earth's magnetic field. Small magnetic sensors such as the ubiquitous *Hall sensor* may be used for many other purposes, such as determining the position or rotational speed of mechanical components. These applications involve **object presence** sensing; see Chapter 3. In that entry, Hall sensors are discussed at “Hall-Effect Sensor”.

In the past, a **magnetometer** was a bulky measurement device incorporating knobs or other controls and some form of display. While that use of the word is still common, this entry deals only with chip-based magnetic sensors.

OTHER RELATED COMPONENTS

- **accelerometer** (see Chapter 10)
- **gyroscope** (see Chapter 9)
- **GPS** (see Chapter 1)

What It Does

A traditional compass consists of a thin magnetized strip of metal balanced on a pivot. It aligns itself with the Earth's magnetic field.

A *scalar magnetometer* measures the total strength of a magnetic field. A *vector magnetometer* can measure the strength in a specified direction. In particular, it can provide a numeric output describing the angle between the orientation of the measuring device and the Earth's magnetic poles.

Chip-based magnetometers are usually vector-type, containing three sensors mounted orthogonally—that is, each of them at 90 degrees to the other two. Suitable software can interpret the analog readings from the sensors to calculate magnetic north or south regardless of the angle at which the instrument is being held, relative to the ground.

Schematic Symbol

There is no specific schematic symbol for a magnetometer.

IMU

A **gyroscope** measures the rate of rotation of the enclosure in which it is mounted. This is properly known as the *angular velocity*. A gyroscope will also respond to changes in the rate of rotation. It does not measure linear motion or a static angle of orientation.

An **accelerometer** measures variations in linear motion and will also measure its own static orientation relative to the force of gravity. If an accelerometer rotates around its own axis, it will not measure angular velocity.

When an accelerometer and a gyroscope are contained in the same package, optionally with a magnetometer, they may be described as an *IMU* (inertial measurement unit), which can pro-

vide necessary data to maneuver aircraft, spacecraft, and watercraft, especially when **GPS** signals are unavailable.

Applications

Magnetometers are found in handheld equipment such as digital compasses, cameras, and cellular phones. They are usually surface-mount chips that are manufactured in large quantities and may be used in conjunction with microcontrollers. For the hobby-electronics community, or for experimental product development, a magnetometer may be mounted on a breakout board for ease of use. A board using a Honeywell HMC5883L is shown in [Figure 2-1](#).

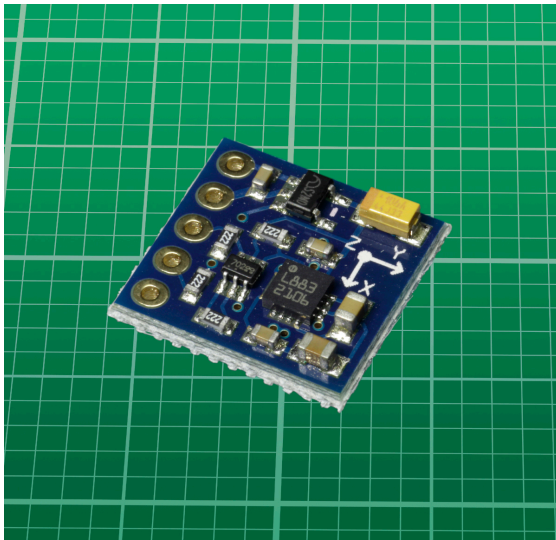


Figure 2-1 The Honeywell HMC5883L 3-axis magnetometer mounted on a breakout board. The background grid is in millimeters.

How It Works

An explanation of magnetometers requires an understanding of the fundamentals of magnetism.

Magnetic Fields

A magnetic field is often represented by *field lines* that show the strength and vector of the field. Field lines associated with a simple per-

manent magnet are shown in [Figure 2-2](#), where the strength, or *flux density*, at any point, is indicated by the spacing the lines, while the angle tangential to a line indicates its vector. (For an extended discussion of magnetism, see **electromagnet** in Volume 1.)

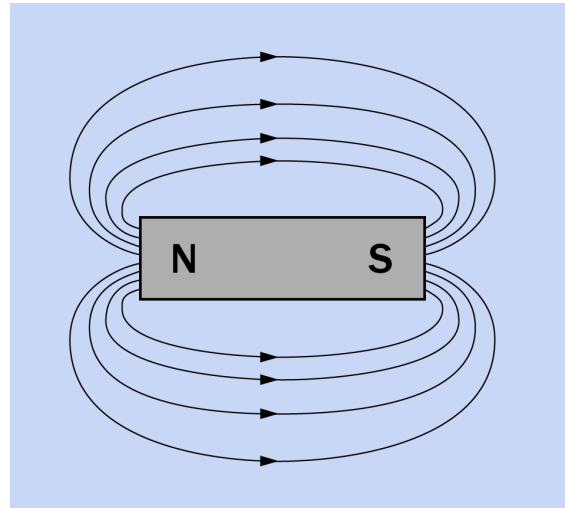


Figure 2-2 Field lines representing magnetic flux created by a bar magnet. Space between the field lines is inversely proportional to flux density. In reality this is a three-dimensional effect, and a more accurate representation would show the magnet and field lines revolved around the axis of the magnet.

The flux density of a magnetic field is usually represented by the letter *B*, and is measured in *newton-meters per ampere*, more commonly referred to as *teslas* (T). An older unit of measurement was gauss (G), 1 tesla being equivalent to 10,000 gauss. Some datasheets still refer to gauss.

The Earth's magnetic field is believed to result from convection currents in the outer liquid of the Earth's core. The strength of this field varies from 25 to 65 microteslas (0.25 to 0.65 gauss) depending on the location where it is measured. To a very rough approximation, the Earth behaves as if a giant bar magnet connects the magnetic north pole and the magnetic south pole. See [Figure 2-3](#).

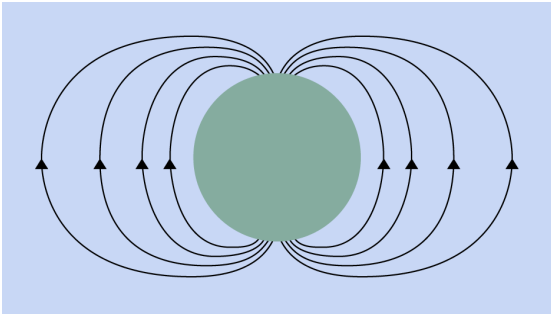


Figure 2-3 The magnetic field of the Earth approximately resembles the field around a bar magnet.

Note that in northern and southern latitudes, the field lines angle steeply down toward the surface, while the lines are approximately parallel with the surface near the equator. Consequently a magnetometer held tangentially to the surface of the Earth will tend to measure a stronger horizontal field near the equator than near the poles.

The varying angle of the field lines tangential to the Earth is known as their *inclination*. Variations in field strength can be used to determine an approximate location, although the **GPS** (Global Positioning System, using satellites) enables this much more accurately.

Confusion results from the fact that the northern magnetic pole of the Earth actually behaves as a south pole, while the southern magnetic pole behaves as a north pole. That is, when a permanent magnet is free to rotate, its north pole will orient itself toward the so-called northern magnetic pole of the Earth, even though opposite poles attract. The northern pole of the Earth should therefore be thought of as the pole that attracts the north end of a compass.

Earth's Axes

Planet Earth revolves around an imaginary line referred to as the *axis of rotation*. This line is close to, but not quite coincident with, the *magnetic axis* connecting the magnetic north pole and magnetic south pole, as shown in Figure 2-4.

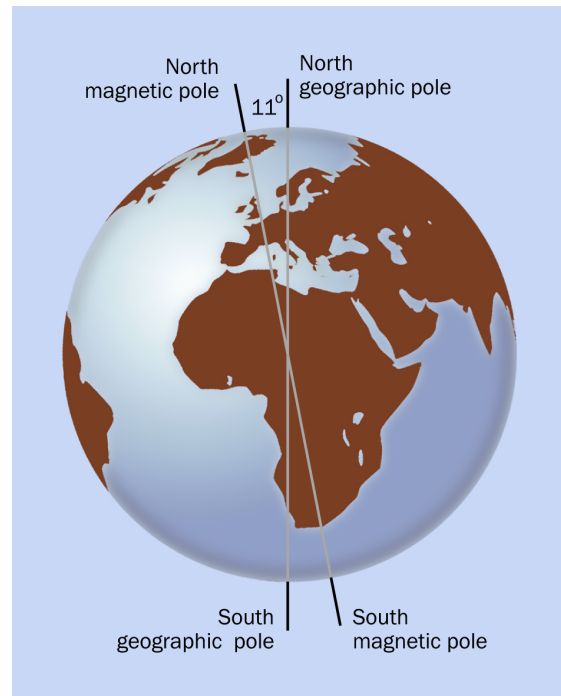


Figure 2-4 The angle between the magnetic axis and the rotational axis of the Earth is approximately 11 degrees.

Magnetic declination is the angle between the magnetic north pole and the geographical north pole, as perceived by an observer. This angle will vary depending on the observer's location on the surface.

Because of declination, the direction of magnetic force at points on the surface of the planet will vary with latitude and longitude, as shown in Figure 2-5. In this figure, the *magnetic meridians* are shown in red, superimposed on the *geographical meridians* shown in green. Magnetic meridians show the direction of magnetic force, while geographic meridians are drawn between the ends of the axis of rotation of the Earth. While there is an approximate correlation between the two, in some areas, especially near the north and south poles, the discrepancy can be more than 40 degrees.

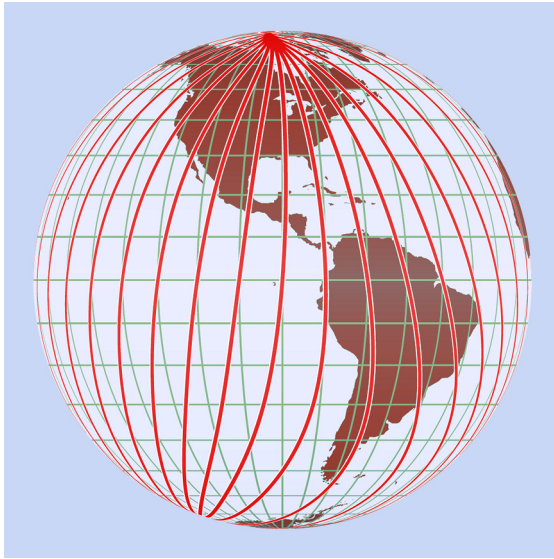


Figure 2-5 The red lines indicate the direction in which a compass would be likely to point to magnetic north. The green lines connect the geographical poles of the planet. (From Wikimedia Commons.)

Standard declination tables for locations on the Earth are available, and these values must be added to, or subtracted from, the reading from a compass or magnetometer to determine the direction of geographical north. Navigational systems customarily express the heading of a vehicle or vessel relative to geographical north, as shown in Figure 2-6.

Coil Magnetometer

Current flowing through a wire creates a magnetic field with flux density that is directly proportional to the current in amperes. Conversely, a changing magnetic field will induce current in a wire. This principle is used in a *coil magnetometer*, which can detect buried objects when the coil moves above them. A *rotating coil magnetometer* can determine magnetic field strength while remaining in a stationary position. However, the coil in a coil magnetometer must be relatively large.

Hall Effect and Magnetoresistance

In modern handheld devices, a magnetometer will generally use either the *Hall effect* (see

"Hall-Effect Sensor") or *magnetoresistance*, described here.



Figure 2-6 Heading is an angle usually calculated relative to geographical north, sometimes referred to as "true north."

Magnetoresistance is a phenomenon where the resistance of a material changes fractionally when it is exposed to a magnetic field. This is capable of yielding greater accuracy than Hall-effect sensors, but has tended to be more expensive.

Orthogonally oriented sensors contained in a surface-mount chip are aligned with axes identified with letters X, Y, and Z. These sensors are analog devices whose values are converted to digital values by an onboard *analog-to-digital converter* (ADC). The values are stored in *registers* that are available to other devices, often via the *I2C* protocol, which is widely used by micro-controllers.

Typically there will be two eight-bit registers for each axis, one defining the high byte and the other defining the low byte for the digital value. In reality the ADC is likely to use 10 to 13 bits, with the remaining 6 to 3 bits being unused.

Variants

The Freescale Semiconductor FXMS3110 is a typical low-cost chip containing a 3-axis magnetometer sensor. Many chips now additionally include a 3-axis accelerometer sensor. An example is the LSM303 by STMicroelectronics, which is sold on a breakout board by Adafruit. See [Figure 2-7](#).

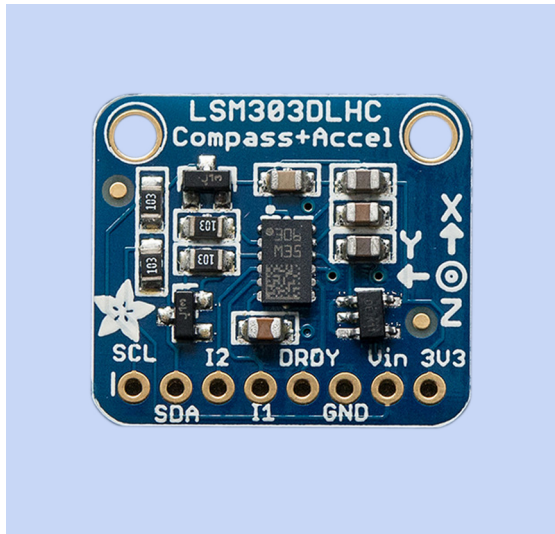


Figure 2-7 The LSM303 is a chip manufactured by STMicroelectronics. It is shown here on a breakout board from Adafruit.

See [Chapter 10](#) for an explanation of the function of accelerometers.

The InvenSense MPU-9250 is a highly sophisticated IMU, including a 3-axis **gyroscope** in addition to a 3-axis magnetometer and 3-axis accelerometer.

A processing unit in the MPU-9250 reconciles the nine variables, and the digital output can be accessed via I2C or SPI protocol at speeds up to 1MHz. All functions of the chip are contained within a package measuring less than 3mm square.

See [Chapter 9](#) for additional details about the functioning of a gyroscope.

How to Use It

A basic 3-axis magnetometer sensor such as the HMC5883L can be tested with a microcontroller that can receive data from its registers via I2C protocol. This is relatively easy with an Arduino, which was designed to be I2C-capable.

Several breakout boards are available with the HMC5883L mounted on them. Many of these boards contain a voltage regulator, allowing a 5VDC power supply to be used even though the chip is designed for a typical supply of 2.5VDC.

In addition to the power supply, the breakout board only requires two connections for I2C communication, to its SCL pin (serial clock input) and SDA pin (serial data input/output). If basic I2C software is installed on the microcontroller, it will read digital values from the magnetometer registers. If additional software is used, it will convert the values to magnetic flux densities in microteslas for the X, Y, and Z axes. Code libraries to achieve these goals are widely available online.

A more sophisticated chip such as the InvenSense MPU-9250 can be used similarly, but yields additional data for conversion. Here again, code libraries can be found online. The slightly older MPU-9150 is sold on a breakout board with downloadable code from Sparkfun.

What Can Go Wrong

Bias

Magnetometers are sensitive to their environment, which can induce *magnetometer bias* of two types.

Hard-iron bias is primarily caused by magnetized material inside the device containing the magnetometer. Because this effect is usually unvarying, compensating for it is relatively easy. *Soft-iron bias* is caused by interaction between variations in the Earth's magnetic field and

materials inside a magnetometer that can be magnetized.

A common example of soft-iron bias would be power lines, generating a magnetic field that can affect model aircraft and drones using magnetometers for navigation.

Mounting Errors

Placement of a chip-based magnetometer on a circuit board is critical. The field effects of trans-

formers or relays must be taken into account, and even the low voltage and low current in a circuit trace can create a magnetic field sufficient to disturb a chip. No traces on any layer of the board should pass across the footprint of the chip. A magnetometer should not be mounted within a ferromagnetic case.

object presence sensor

An **object presence** sensor may also be described as an *object detector* or *detection sensor*.

The term **proximity** sensor may be applied to this component. However, in this Encyclopedia a proximity sensor has the capability of estimating the distance to a target. See Chapter 5. An object presence sensor merely detects whether an object is within a preset range, and does not provide additional information.

Optical and *magnetic* object-presence sensors are described and compared in this entry. *Ultrasonic sensors* are described in the entry dealing with proximity sensors, because they tend to be used for distance measurement rather than just for detection. Other methods of object presence sensing, including *capacitive*, *doppler*, *inductive*, *radar*, and *sonar*, are not included in this Encyclopedia.

A sensor that detects an object by receiving light reflected from it is categorized as a *reflective* sensor, and is included in this entry. (If a module includes a light source as well as a light sensor, it is properly categorized as *retroreflective*, although that term is not always used.)

A sensor that detects an object when it interrupts a beam of light is a *transmissive* sensor, and is included in this entry. It may also be described as a *through-beam* sensor or as an *optical switch*.

A sensor that responds to the motion of an object that emits infrared radiation is a **passive infrared** motion sensor, also described by its acronym, PIR, and often referred to simply as a *motion sensor*. It has its own separate entry. See Chapter 4.

Phototransistors and **photodiodes** may be used as sensing elements in presence sensors. These components have their own entries as light sensors. See Chapter 22 and Chapter 21.

OTHER RELATED COMPONENTS

- **proximity** sensor (see [Chapter 5](#))
- **passive infrared** sensor (see [Chapter 4](#))

What It Does

An **object presence** sensor verifies the presence or absence of an object within a predeter-

mined range, without necessarily measuring how far away it is or how fast it is moving. The object may be described as a *target*.

Object presence sensing is often used to verify the correct function of an automated system—for example, the placement of objects on a conveyor belt. It can also be used to count objects as they pass a sensor.

Some types of security systems use presence sensors to sound an alarm if an intruder interrupts a beam of light. They can verify that a door or window is closed. They may also function as a limit switch to control the operation of a motor.

Schematic Symbol

In schematics, an optical presence sensor may be indicated with the symbol for an LED, plus the symbol for a **phototransistor**, with one or two arrows connecting them, as shown at top-left in Figure 3-1. Wavy-line arrows may indicate an infrared connection.

A **photodiode** may be substituted for a phototransistor, as shown at top-right in the figure.

A magnetic sensor may be shown with the symbol for a *Hall-effect sensor*, as in the lower half of Figure 3-1.

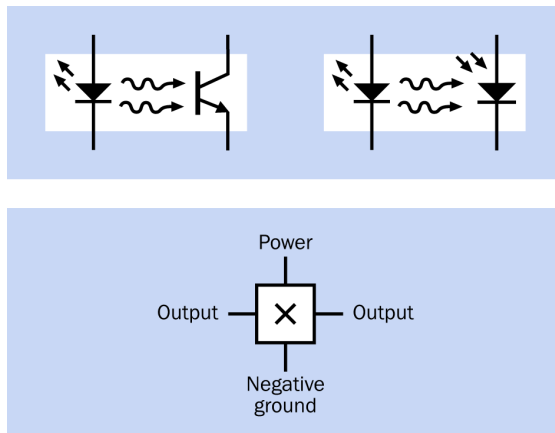


Figure 3-1 Top: two possible schematic symbols for an optical presence sensor, using a phototransistor (left) or photodiode (right). Other variants are possible. Bottom: the schematic symbol for a Hall-effect sensor, commonly used in a magnetic sensor.

Variants

To assist the reader in comparing different options for detecting the presence of an object, this entry includes two primary variants: *optical* and *magnetic*.

The optical sensors are divided into *transmissive* and *reflective* (including *retroreflective*). The magnetic sensors are *reed switches* and *Hall-effect sensors*. A chart showing the categories and subcategories is shown in Figure 3-2.

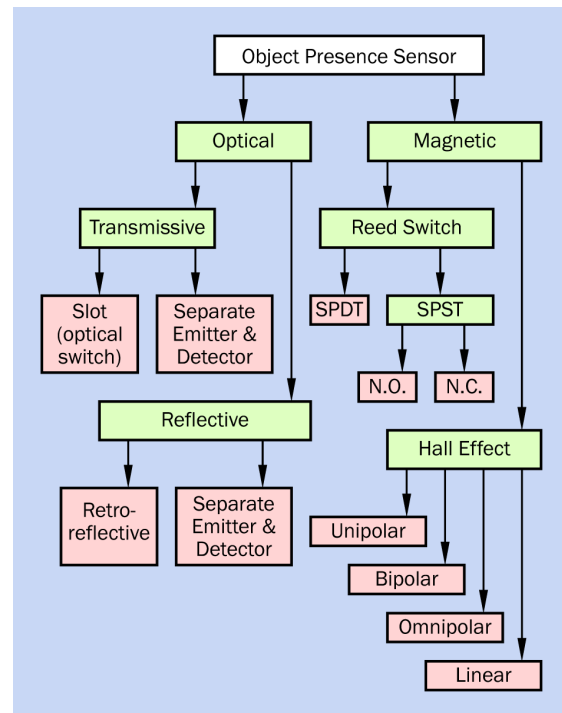


Figure 3-2 Categories of object presence sensors discussed in this entry. Other types of object presence sensors exist, but are less common, and are not included here.

Optical Detection

A *transmissive optical sensor*, also known as a *through-beam sensor*, is really a pair of components, one emitting light and the other receiving it. The sensor is triggered if an object

interrupts or reflects the light beam, as shown in [Figure 3-3](#).

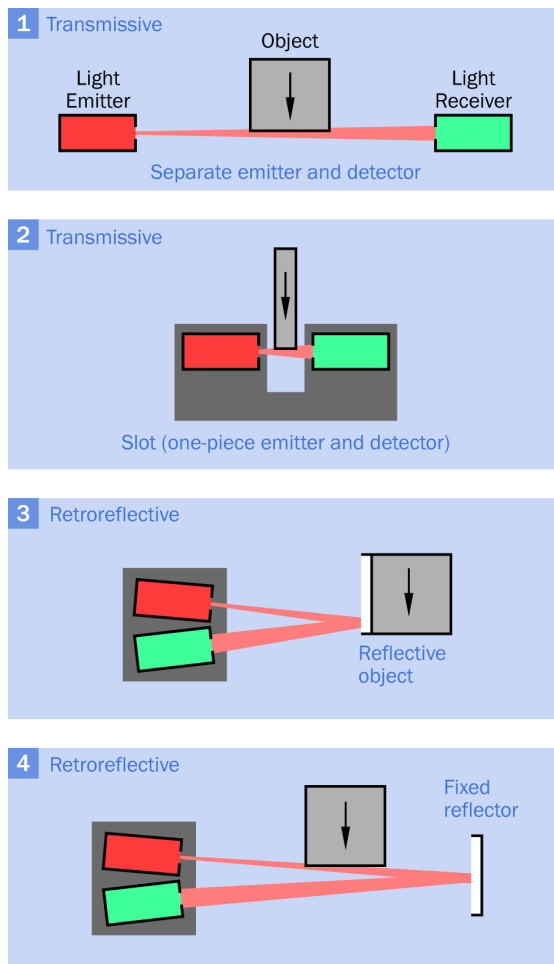


Figure 3-3 Various configurations of optically activated object presence sensors. See text for details.

If the light emitter and the light detector face each other across a small gap, they may be contained in a single module (usually with a slot in it) as shown in section 2 of [Figure 3-3](#). This is often referred to as an [optical switch](#) (not to be confused with the solid-state switching devices used in telecommunications). It is sometimes described as a [photointerruptor](#) or [photointerrupter](#).

A [reflective optical sensor](#) also consists of a light emitter and a light detector, but they are

placed adjacent to each other, facing in the same general direction. When they are mounted in one module, as is often the case, the arrangement is properly known as a [retroreflective sensor](#). Either way, the combination may be triggered in one of two ways:

- An object passing in front of the light beam reflects it back to the detector. The object must be naturally reflective (for example, glass containers or white boxes on a conveyor), or must have a reflective patch applied to it, or the light source must be bright enough to reflect from an object that is not highly reflective. This configuration is shown in section 3 of [Figure 3-3](#).
- A stationary reflector can be mounted opposite the light emitter, in which case a detector beside the light emitter is triggered when an object interrupts the reflected light beam. This configuration is shown in section 4 of [Figure 3-3](#).

Transmissive Optical Sensors

A light source and light detector may be sold as separate components in a matched pair. An example is the Vishay TCZT8020, shown in [Figure 3-4](#). These components are small, each measuring no more than 5mm x 3mm. They are designed to be placed just a few millimeters apart. The light source is an infrared LED, while the detector is a **phototransistor** (see [Chapter 22](#) for information about phototransistors).

The source and detector are both designed to use 5VDC. Output from the phototransistor is an open collector. Current through the collector must not exceed 50mA, and must be limited by a pullup resistor of 100 ohms or more. Current through the source must not exceed 60mA, and must be limited by an appropriate series resistor.

Details on using an open-collector output will be found in the Appendix. See [Figure A-4](#).

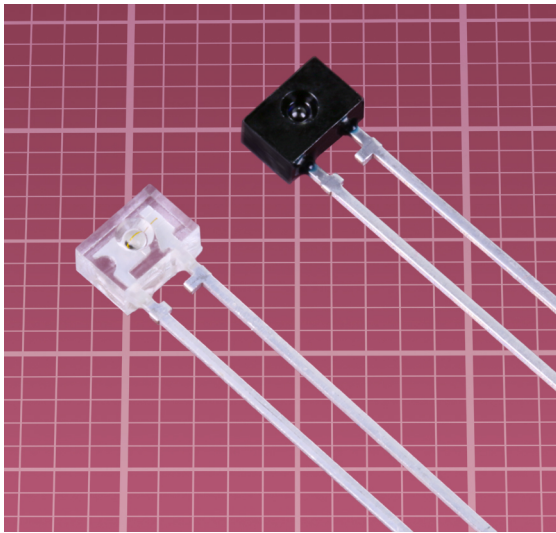


Figure 3-4 A matched light source and light detector, for use as a transmissive optical sensor. The background grid is in millimeters.

The Omron EE-SX range includes a variety of configurations of light source and detector separated by a 5mm slot in one module. The source is an infrared LED, and the detector is a phototransistor.

The Omron components tolerate a wide range of supply voltages, from 5VDC to 24VDC, with no series resistor necessary for the LED. Open-collector output from the phototransistor can tolerate 50mA to 100mA, depending on the particular version of the sensor. A red indicator LED is illuminated when an object blocks the slot in the sensor. Some versions have a high output when the slot is open, while others have a high output when the slot is closed. Because of their various features, these sensors are relatively expensive.

A much cheaper optical switch is the Everlight ITR9606 (described by the manufacturer as an “opto interrupter”). It is pictured in [Figure 3-5](#). This is intended as a 5V device and has an open-collector output. It requires a series resistor for the LED and a pullup resistor for the open-collector output. Many similar detectors are available.

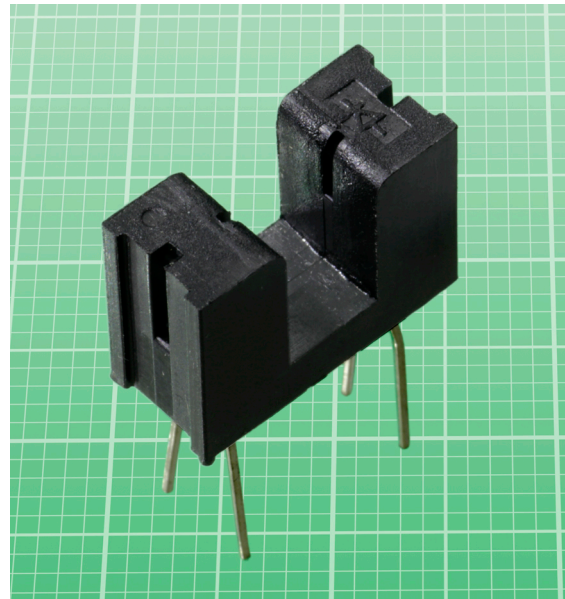


Figure 3-5 A low-cost transmissive sensor, more commonly known as an optical switch. The background grid is in millimeters. The slot in the sensor is about 5mm wide.

For longer-distance detection, an infrared receiver can be mounted separately from an infrared LED. The TSSP77038 from Vishay detects infrared light from as far away as 50cm, and pulls its open collector output low in response. The light must be modulated at 38kHz.

Polulo Robotics and Electronics sells a very affordable breakout board containing a TSSP77038 receiver paired with an infrared LED modulated by a 555 timer. It is shown in [Figure 3-6](#). Because this board contains a light source as well as a light detector, it is really a retroreflective sensor.

Where distances exceeding 1 meter are involved, a laser coupled with a phototransistor that is shielded from ambient light may be necessary.

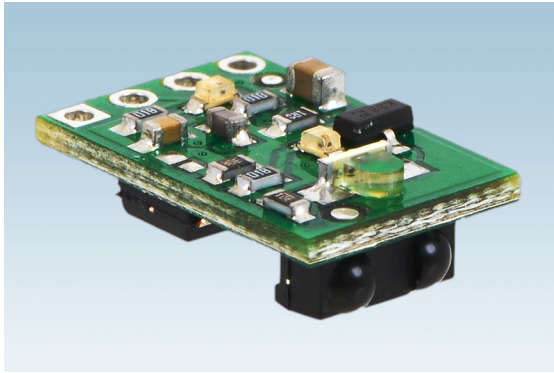


Figure 3-6 The Vishay TSSP77038 installed with an appropriate light source on a breakout board from Polulo Robotics and Electronics.

Retroreflective Optical Sensors

As is the case with a transmissive object detector, the retroreflective type may be listed by vendors as an *optical switch*. Other terms used in datasheets include *reflective interrupter*, *reflective object sensor*, *reflective optical sensor*, *reflective photointerrupter*, *opt-pass sensor*, and *photomicrosensor (reflective)*. The remarkable lack of standardization in terminology creates a problem when searching online for these devices. Why so many different names evolved for the exact same component is unclear.

Many retroreflective object-detection sensors are available in packages ranging from 5mm x 5mm to 10mm x 10mm in size. Almost all of these modules are analog devices using an infrared LED as the light source and a phototransistor as the sensor, with an open-collector output. (For more information about phototransistors, see [Chapter 22](#).)

With a suitable pullup resistor, the output voltage will be proportional with the inverse of the distance. If V is the voltage, d is the distance, and k is a conversion factor:

$$V = k * (1 / d)$$

While many of the smaller modules are surface-mount, some have leads, as shown in [Figure 3-7](#). A major limitation of these small components is that they have a very limited

range, typically less than 5mm. They can only be used in applications such as process control where the position of a target is controlled and predictable.

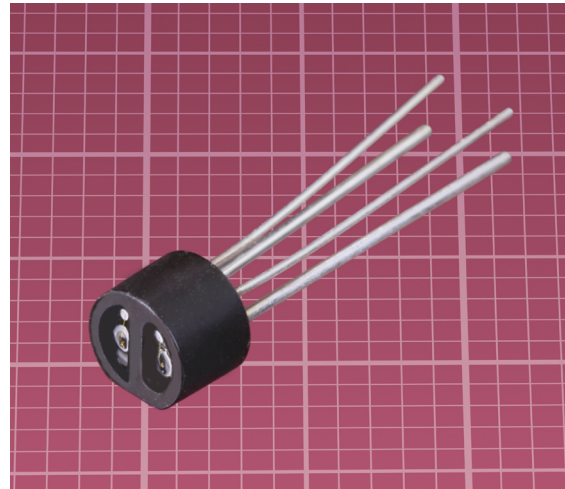


Figure 3-7 The Rodan RT-530 is a small object-presence sensor with a limited range typical of this type of retroreflective component. The background grid is in millimeters.

Another example of a retroreflective sensor in a small package is the Optek OPB606A shown in [Figure 3-8](#). The background grid is in millimeters.

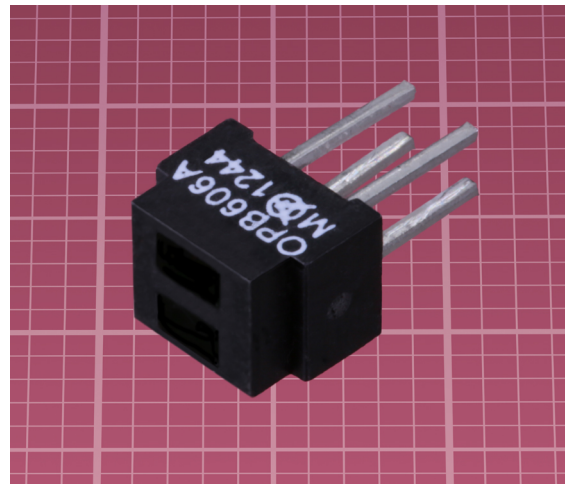


Figure 3-8 The Optek OPB606A. The background scale is in millimeters.

A retroreflective module, with a lensed LED and lensed phototransistor to focus the outgoing and reflected beams, is the Vishay TCRT5000 shown in Figure 3-9.

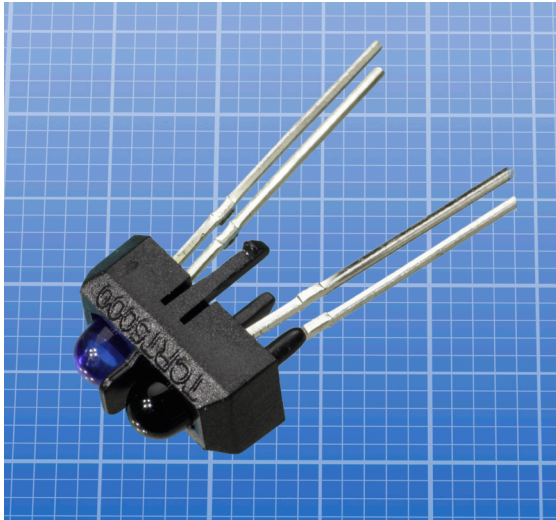


Figure 3-9 The Vishay TCRT5000 retroreflective sensor. The background grid is in millimeters.

Retroreflective modules that have a greater range tend to be physically larger, less common, and more expensive. Sharp makes a popular series. Some part numbers and distance sensing limits are: GP2Y0D805Z0F (5mm to 5cm), GP2Y0D810Z0F (2cm to 10cm), and GP2Y0D815Z0F (5mm to 15cm). Figure 3-10 shows the GP2Y0D810Z0F mounted on a small breakout board from Polulo Robotics and Electronics. The board is useful, as the pins on the sensor are spaced only 1.5mm apart. The size of the board is approximately 8mm x 20mm.

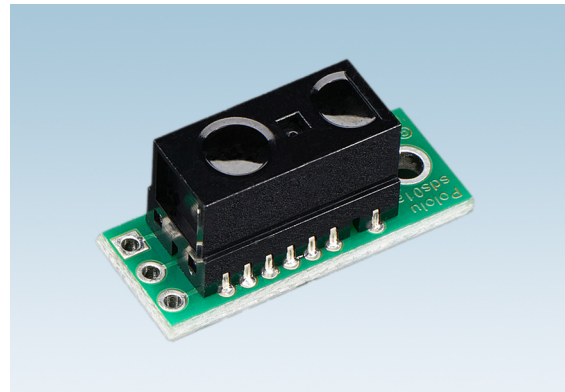


Figure 3-10 The Sharp GP2Y0D810Z0F detection sensor mounted on a board from Polulo Robotics and Electronics. From a photograph by Adafruit Industries.

Each of the detection sensors in this series is described in Sharp datasheets as a “Distance Measuring Sensor Unit,” but in fact they do not measure distance. A single output is normally logic-high and drops to logic-low when a target is within range. Sharp refers to this as a “digital” output, but in fact it is binary, and should not be confused with the digital buffer on a more sophisticated proximity sensor that contains an analog-to-digital converter and provides a numeric output.

It is important to distinguish the Sharp object-presence sensors listed above from the range of Sharp **proximity** sensors described in the entry in Chapter 5. The proximity sensors are physically larger, and most have an analog output that varies with the distance of a target.

Magnetic Sensors

Prepackaged magnetic sensing units are sold in many configurations for industrial and military applications. Although they may be referred to as “magnetic sensors,” they are outside the scope of this Encyclopedia. Here we discuss board-mounted components. Almost always, they use a *reed switch* or a *Hall-effect* sensor as their sensing elements.

Reed Switch

A reed switch is a magnetically activated mechanical switch. It consists of two metallic contacts in a small enclosure that is usually a glass capsule. The contacts are magnetic, and move in response to a magnetic field. A permanent magnet is used to activate the switch. Two reed switches are shown in [Figure 3-11](#).

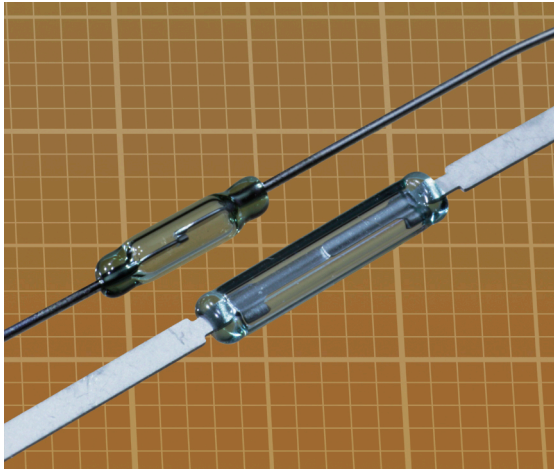


Figure 3-11 Two SPST reed switches. Although the contacts may seem to be touching, in fact they are separated by a tiny gap, and these switches are the normally open type. The background grid is in millimeters.

A reed switch exhibits a small amount of *hysteresis*, because the magnetic field strength required to overcome the mechanical resistance of the springy contacts is greater than the field strength required to keep them closed.

Very small electromagnetic relays that only switch a very low current may use a coil-activated reed switch. For purposes of this Encyclopedia, such a component is considered to be a relay, not a sensor. For more information about **relays**, see the entry in Volume 1.

The most common everyday application for a reed switch is in an alarm system that is triggered by an intruder entering a building. A magnet in a sealed plastic enclosure is attached to a door or window, while a reed switch, in another sealed plastic enclosure, is attached to

the frame, very close to the magnet. Typical components of this type are shown in [Figure 3-12](#). A diagram illustrating the mode of operation appears in [Figure 3-13](#).



Figure 3-12 A typical alarm sensor to detect the opening of a door or window. The nearer module contains a magnet, while the module behind it contains a reed switch.

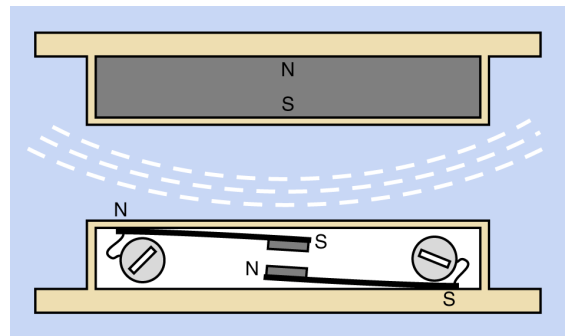


Figure 3-13 The dashed white lines suggest the magnetic field that closes the contacts of a reed switch.

So long as the alarmed door or window remains closed, the magnet activates the reed switch. If the door or window is opened, the magnet moves away from the reed switch, and its contacts relax. Usually in this application the reed switch is the normally open type, and is held in its closed state by the magnet. This allows multiple switches to be wired in series, completing a circuit. If a single switch opens, the circuit is broken, and an alarm is triggered.

Reed Switch Variants

Most reed switches are SPST, either normally open or normally closed, although normally

open variants are more common. Some switches are SPDT, although this variant is relatively rare. An example of a SPDT reed switch is shown in Figure 3-14.

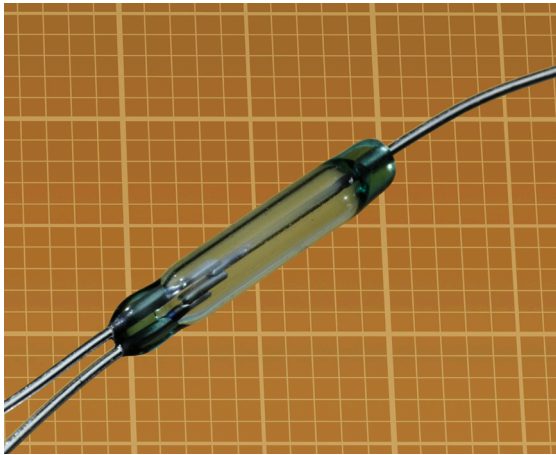


Figure 3-14 A SPDT reed switch. The background grid is in millimeters.

The physical size of a reed switch tends to vary roughly in proportion with its current-switching capability.

Reed switches are most commonly supplied with axial leads. A minority are available for surface-mount applications.

Some reed switches are packaged in a plastic pod to provide physical protection for the glass capsule.

Reed Switch Values

The datasheet for a reed switch is likely to contain the following values:

Pull-in: The minimum magnetic field strength required to activate the switch, often measured in ampere-turns.

Drop-out: The maximum magnetic field strength that allows the contacts of the switch to relax, often measured in ampere-turns. Pull-in will be a higher value than drop-out.

Maximum switching current: While a few industrial reed switches can switch as much as 100A, they are uncommon and expensive. A common

value for a reed switch about 15mm long is 500mA.

Maximum carry current: If specified, it will be higher than the switching current.

Maximum switched power: Because a reed switch can be used with alternating current, its switching capability may be expressed in watts, or as VA (volts multiplied by amps). 10VA is a common value.

Maximum voltage: While reed switches are most often used with low voltages, some are rated to switch up to 200V.

How to Use a Reed Switch

While an optical object presence sensor may be supplied in a single package containing a light emitter and a light detector, a reed switch always requires an activating magnet that is mounted separately. For reliable operation, the maximum distance between the switch and the magnet is usually restricted to a few millimeters.

The orientation of a magnet that activates a reed switch is not crucial, but will affect the sensitivity of the switch. A manufacturer's datasheet should be consulted for information about optimal magnetic polarity.

As in other mechanical switches, the contacts in reed switches vibrate very briefly when opening or closing. This is known as **contact bounce**, and can be misinterpreted as a series of separate signals by digital logic or a microcontroller. **Debouncing** with hardware designed for that purpose, or software (in a microcontroller), may be necessary. See the entry on **switches** in Volume 1.

Hall-Effect Sensor

A Hall-effect sensor responds to a magnetic field by generating a small voltage, usually amplified by a transistor included in the package with the sensor.

When a Hall sensor is in its “off” state (i.e., not triggered by a magnetic field), the sensor creates high resistance between the collector of an internal NPN transistor and negative ground. Consequently, if voltage through a pullup resistor is applied to the collector, the voltage on the collector will be high.

When the sensor is in its “on” state, the resistance drops, the voltage supplied to the collector from the pullup resistor is shunted to ground, and the output voltage appears to go low. As a general rule:

- Activated Hall sensors appear to have a low output.
- Inactive Hall sensors appear to have a high output.

Information about using open-collector outputs is contained in the Appendix. See [Figure A-4](#).

The clean output, reliability, small size, and cheapness of Hall sensors have encouraged their use in devices as different such as hard drives, cameras, keyboards, and automobiles. They are useful in almost any situation where a sensor must detect a mechanical operation at close range. Three through-hole Hall sensors are shown in [Figure 3-15](#). Surface-mount versions are much smaller.

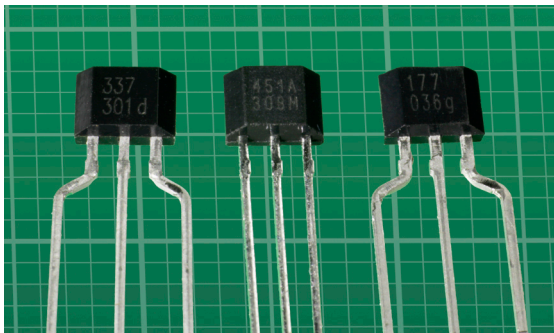


Figure 3-15 Three through-hole Hall-effect sensors. The background grid is in millimeters.

How a Hall-Effect Sensor Works

When current is flowing through the length of a conductor, and a magnetic field is applied across the width of the conductor, the field generates a force causing electrons and electron holes to accumulate asymmetrically on opposite sides. This is known as the *Hall effect*.

The voltage difference between the electron-rich and electron-depleted zones is known as the *Hall voltage*. It is proportional to the magnetic field, and inversely proportional to the density of free electrons in the material. For this reason, the Hall effect is easiest to observe in semiconductors, where the density of electrons or electron holes is low.

Hall sensor components contain amplifier circuitry in addition to the sensing element itself. Typically there is an open-collector output, and a comparator or Schmitt trigger to provide some hysteresis.

Hall-Effect Sensor Variants

Four primary variants of Hall sensors are widely used.

Unipolar Hall Sensor

This is activated when an external magnetic field exceeds a threshold value. When the field diminishes, the sensor switches off. Unipolar sensors are available in versions activated by the north magnetic pole or south magnetic pole.

Bipolar Hall Sensor

Proximity to one magnetic pole will switch it on. Proximity to the opposite magnetic pole will switch it off. The sensor remains in its current state (on or off) in the absence of a magnetic field.

Omnipolar Hall Sensor

Proximity to a strong magnetic field of either polarity will switch it on. Removal of the magnetic field will switch it off. An omnipolar sensor can be thought of as a pair of unipolar sensors mounted in opposite directions and

with their (open-collector) outputs wired together. This component functions similarly to a reed switch, although of course it still requires a power supply.

Linear Hall Sensor

Also known as an *analog* Hall sensor, its output voltage varies in proportion to an external magnetic field instead of switching cleanly between high and low states. When no magnetic field is detected, the output is half of the sensor's supply voltage. In response to one magnetic polarity, the output can drop almost to zero. The opposite polarity can increase the output almost to supply voltage.

The output from a linear sensor usually is supplied from the emitter of an internal NPN transistor, not the collector. A minimum 2.2K resistor should be connected between output and ground.

The variable output can be interpreted as a measurement of distance between the sensor and a magnet. In this mode, a Hall sensor functions as a **proximity** sensor. However, it is not usually capable of measuring a distance of more than 10mm.

Other Applications

Hall sensors are incorporated in other types of components. A **magnetometer**, for example, may contain Hall sensors.

Additional discussion of Hall-effect sensors, with test circuits, will be found in the book *Make: More Electronics*, from which some of the figures here have been excerpted.

Values

Magnetic field at operating point is the minimum field necessary for the output to switch on. It is measured in tesla or gauss, and the abbreviation B_{OP} is used.

Magnetic field at release point is the maximum field that allows the output to switch off. It is

measured in tesla or gauss, and the abbreviation B_{RP} is used.

Magnetic field range may be specified for linear (analog) Hall sensors.

Supply voltage may range as widely as 3VDC to 20VDC, or may be restricted between 3VDC to 5.5VDC. Check datasheets carefully.

Sourcing or sinking capability for the open-collector output is typically 20mA.

How to Use a Hall-Effect Sensor

Hall sensors are often made in 3-pin packages. Through-hole variants are usually made of black plastic and look like TO-92 transistors, but are slightly smaller.

Surface-mounted variants are common.

A typical through-hole Hall sensor has one bevelled face and a flat face on the opposite side. The bevelled face may be referred to as the "front" of the component in a datasheet. The sensor responds when an appropriate magnetic pole is brought close to the front face of the sensor.

The part number printed on the front of the sensor may be abbreviated as three digits. The code below this usually refers to the date of manufacture.

A simple circuit for a Hall sensor resembles a typical circuit for a phototransistor. Positive supply voltage and negative ground are applied to two of the three leads. Positive voltage is also applied, through a pullup resistor, to the third lead, which is the open-collector output (except in the case of a linear Hall sensor, described previously). The output pin is then tapped as the output from the sensor, to be applied to a component that will not draw more than 20mA.

Configuration of Object Presence Sensors

While most of the following suggestions relate specifically to Hall-effect sensors, some general principles may be applied to optical sensors.

Linear Motion

A presence sensor can be activated when the triggering source (such as light or a magnet) approaches it directly. This is sometimes referred to as *head-on mode*. Alternatively, triggering can be arranged when the source moves past the sensor. This is sometimes referred to as *slide-by mode*. The two modes are illustrated in sections 1 and 2 of Figure 3-16.

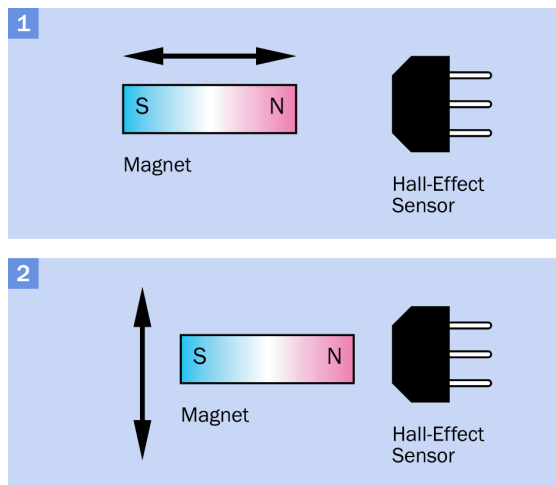


Figure 3-16 Section 1 of this figure illustrates a presence sensor being used in head-on mode, while section 2 illustrates slide-by mode.

Slide-by mode may be preferred because it eliminates the risk of damage to the sensor if overshoot occurs in head-on mode.

In slide-by mode, using a bipolar Hall-effect sensor, two magnets can be placed together with opposite polarity, creating a very steep transition in the overall magnetic field. This minimizes the risk of imprecise triggering. Using neodymium magnets, the triggering

point can be adjusted with a precision of 0.01mm or better. See Figure 3-17.

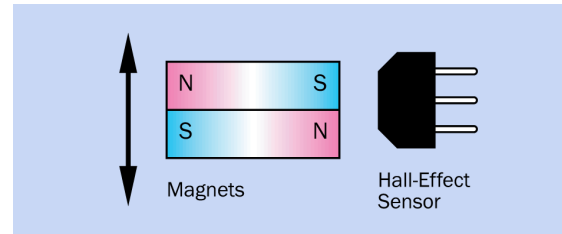


Figure 3-17 In slide-by mode, two magnets with opposite polarity can be put together to create a very precise transition in a bipolar Hall-effect sensor.

Sensing by Interruption

An *optointerrupter* is sensitive to an object passing between the light source and the light sensor. A comparable arrangement can be used with a Hall-effect sensor or a reed switch, but only if the interrupting object is thin and ferrous. This configuration is known as a *ferrous vane interruptor*.

Note that the magnet will exert significant force on the ferrous vane. This becomes an issue in sensors where the mechanical force is limited—for example, in paper-path sensors in a photocopy machine.

Additional information about the detection and measurement of moving objects will be found in the entry describing **linear position** sensors. See Chapter 6.

Angular Motion

One or more magnets can be used with a Hall-effect sensor to detect the angular motion, relative angular position, or absolute angular position of a rotating part. This data can be used to determine its speed of rotation. Some techniques to achieve this are described in the entry describing **rotary position** sensors. See Chapter 7.

Sensor Comparisons

Advantages of Optical Presence Sensors

- Not affected significantly by magnetic fields that can interfere with operation of a Hall-effect sensor or a reed switch.
- May be contained in a small, all-in-one package.
- Some optical sensors can operate over a distance of 50cm.
- Very well suited to sensing an object that blocks the light source (the *optointerrupter* configuration).

Disadvantages of Optical Presence Sensors

- Must have clear line-of-sight with the object and/or a reflector.
- Performance will be degraded by accumulation of dust or dirt.
- Limited lifetime of LED light source, if used continuously.
- May be accidentally triggered or impaired by some types of ambient light.
- Often requires a load resistor for the LED as well as a pullup resistor for the open-collector output.
- Range of acceptable voltage is usually narrow.

Advantages of a Reed Switch

- No polarity.
- No additional components required, other than a magnet.
- Able to switch AC or DC.
- Able to switch to voltages as high as 200V in some cases.

- Can be maintained by a magnet in open or closed state indefinitely without any power consumption.
- Many variants are capable of switching 500mA, and higher-current types are available.
- Can be activated through nonmagnetic materials (plastic, paper).
- Not significantly affected by dust or dirt that can degrade performance of an optical switch.

Disadvantages of a Reed Switch

- Requires a separate magnet (which must be placed carefully to avoid affecting other components).
- Cannot be miniaturized to the same extent as a surface-mount chip.
- Glass envelope is easily damaged.
- Arcing may occur between contacts.
- Will not work reliably when the activating magnet is more than a few millimeters from the switch.
- Can be activated accidentally by other magnetic fields.
- When sensing an object that comes between the switch and a magnet, only a ferrous object can be used.
- Must be debounced when connected with a logic chip or controller.

Advantages of a Hall Effect Sensor

- Robust solid-state component.
- Can be miniaturized for surface-mount applications.
- Very low cost.
- Fast response.
- No contact bounce.

- Extremely durable, with an almost unlimited lifetime.
- Not significantly affected by dust or dirt that can degrade performance of an optical switch.
- Low cost.

Disadvantages of a Hall Effect Sensor

- Requires a separate magnet (which must be placed carefully to avoid affecting other components).
- Open-collector output typically limited to 20mA or less.
- May be vulnerable to magnetic fields.
- When sensing an object that comes between the switch and a magnet, only a ferrous object can be used.

What Can Go Wrong

Optical Sensor Issues

Deterioration of LEDs

Most object presence sensors depend on infrared LEDs as a light source. An LED has many virtues (discussed in Volume 2), but does tend to suffer from gradual reduction in light output over a period of years. In a device such as a photocopy machine, which is used intermittently and may have an “economy” or “sleep” mode in which most of its components are powered down, LED-based detectors should last almost indefinitely. In other applications where an LED is powered continuously, its light output will diminish significantly in 3 to 5 years. Bearing this in mind, an optical sensor should be chosen to operate well below its light-detection limits.

Object Too Close

Some optical and ultrasonic detectors triangulate on an object; that is, the light emitter and

light sensor are angled slightly toward each other (as shown in sections 3 and 4 of [Figure 3-3](#)). The signal output from the sensor will peak at the distance where the emitter and sensor focus at a point. Consequently the output voltage will diminish when the object comes closer, which may create the misleading impression that the object is moving further away. To avoid spurious readings, detectors should not be used with objects that are closer than the minimum distance specified by the manufacturer.

Reed Switch Issues

Mechanical Damage

Bending the axial leads on a reed switch can easily fracture the glass envelope. Reed switches must be handled with care.

Contact Bounce

If the switch is wired to the input of a logic chip or microcontroller, contact bounce when the switch opens or closes is likely to be misinterpreted as multiple switching events. Debouncing will be necessary, either by additional components or by insertion of a momentary delay in the code embedded in a microcontroller.

Arcing

When switching high voltages or currents, an arc may be created briefly between the switch contacts, most often when they are moving from the closed to the open state. The arc erodes the contacts of the switch. Inductive loads make the arcing problem worse. If the switched voltage is kept below 5V arcing generally does not occur, extending the life of the switch.

passive infrared sensor

4

In colloquial speech, the term *motion sensor* is usually understood to mean a **passive infrared** motion sensor.

The acronym *PIR* is often used for a passive infrared sensor. It is always capitalized, without periods.

Object presence sensors and **proximity** sensors require an active source of a magnetic field, ultrasound, or infrared radiation. A passive infrared sensor does not require any such source, and responds passively to heat radiated from the object that is being detected.

OTHER RELATED COMPONENTS

- **object presence** sensor (see [Chapter 3](#))
- **proximity** sensor (see [Chapter 5](#))

What It Does

A **passive infrared** motion sensor, often described as a *PIR*, detects *black-body radiation*, which all objects emit as a function of their temperature relative to absolute zero. The sensor responds to infrared radiation centered around a wavelength of 10 μ m (10 microns, or 10,000nm). This is the approximate body temperature of people and animals.

The word “passive” in the term “passive infrared” refers to the behavior of the detector, which receives infrared radiation passively. **Proximity** sensors must generate their own infrared radiation actively, which is interrupted or reflected by nearby objects. See [Chapter 5](#).

Schematic Symbols

Schematic symbols that are sometimes used to represent a passive infrared motion sensor are shown in [Figure 4-1](#).

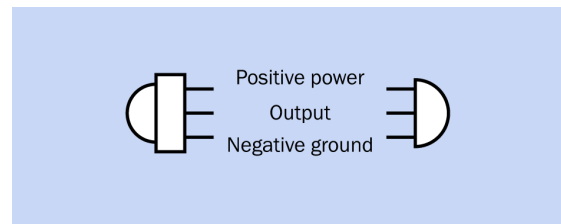


Figure 4-1 Schematic symbols that may represent a passive infrared motion sensor. The orientation (pointing left or right) is arbitrary. The pin sequence may vary.

Applications

Motion-sensitive outdoor lighting almost always is based around a PIR. Similarly, a security system may sound an alarm or activate a video camera when a PIR indicates human activity.

Wildlife monitoring systems use PIRs to start a video camera that can then run for a preset interval.

Warning systems for automobiles have been developed that use a rear-facing PIR to detect pedestrians.

Industrial indoor lighting may use PIRs that switch the lights on automatically when people enter a room, and then switch the lights off (after a timed delay) when people are no longer detected in the room. The goal is to prevent wastage of energy as a result of employees forgetting to switch the lights off.

How It Works

A PIR module contains multiple components. Most visible is an array of at least 15 small lenses that focus infrared light from zones in the environment onto a *pyroelectric detector*, also known as a *pyroelectric sensor*. The response of the detector is processed by an amplifier, so that the signal can trigger an electromechanical **relay** or **solid-state relay** (see Volume 2). The relay operates an external device such as a light or an alarm.

Additional circuitry may allow the user to control the sensitivity of the PIR module and the length of time that the relay remains closed. The user may also be able to set the time of day when the PIR is active, or an additional **photo-transistor** can shut down the PIR during daylight hours. If a phototransistor is included, its sensitivity is adjustable. For more information about phototransistors, see [Chapter 22](#).

Pyroelectric Detector

The pyroelectric detector is actually a type of *piezoelectric* device. It is based around a wafer of lithium tantalate, which generates a small voltage in response to incident thermal radiation. However, like other piezoelectric components, it does not respond to a steady-state input, and must be activated by a transition. This distinguishes it from other types of light sensors, such as an infrared **photodiode**, in which the response is consistently related to a temperature input.

The response of a pyroelectric detector is suggested by the graphs in [Figure 4-2](#).

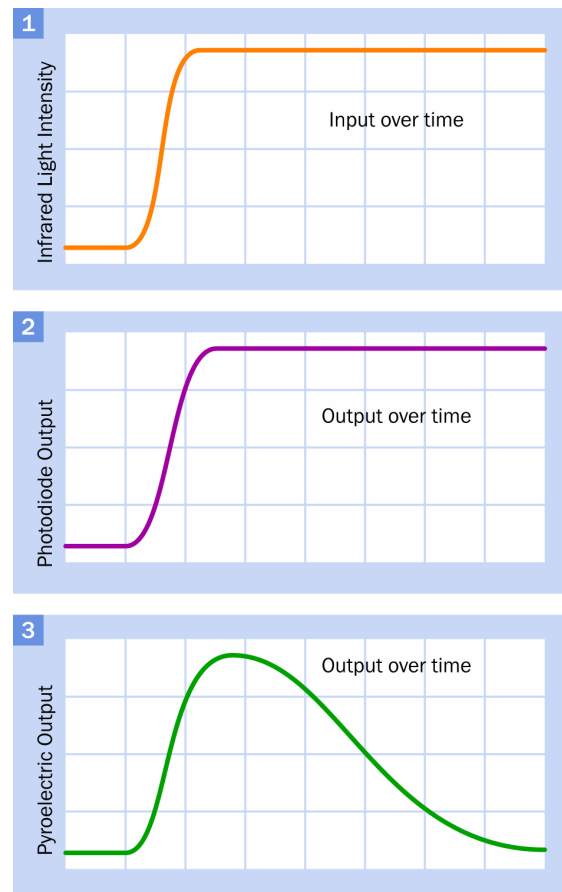


Figure 4-2 Top: incident infrared light intensity. Center: the voltage output from a hypothetical photodiode. Bottom: the voltage output from a hypothetical pyroelectric detector.

A pyroelectric detector in a PIR module is mounted in a sealed metallic container, as shown in [Figure 4-3](#). The rectangular window in the detector is usually made of silicon, which is opaque to visible wavelengths but transparent to long-wave infrared radiation.

Elements

The pyroelectric detector used in a PIR contains at least two *elements* with opposite polarities, connected in series. If a sudden change in temperature affects both elements equally, their

responses will cancel each other out. Thus, the detector ignores changes in ambient temperature. However, if a source of infrared radiation in the appropriate waveband affects one element before the other, the detector will emit two pulses of opposite polarity. See [Figure 4-4](#).



Figure 4-3 A pyroelectric detector mounted on a small circuit board in a passive infrared radiation sensor.

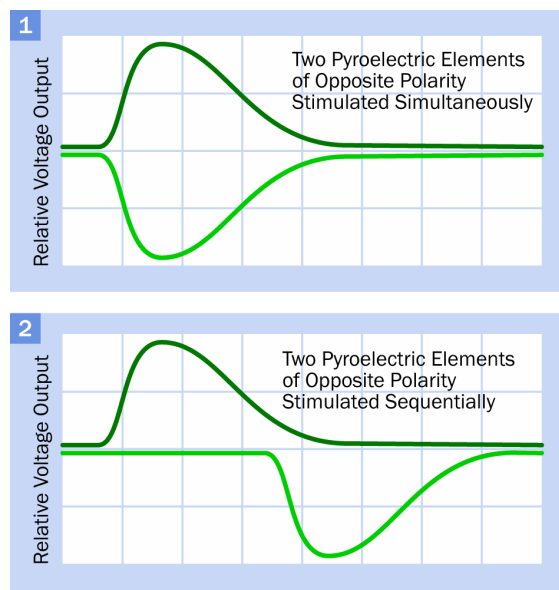


Figure 4-4 Top: in a pyroelectric detector, if a change in temperature affects two elements of opposite polarity simultaneously, their voltages cancel each other out. Bottom: if one element is triggered before the other, the detector emits a signal.

Lenses

Lenses are used to trigger the elements sequentially. Each lens faces toward one visible zone in the target area. When a source of infrared radiation moves from one zone to the next, it energizes the elements alternately, creating an output.

In some PIRs, four pyroelectric elements are used instead of two, to provide better coverage. The pairs of sensors may be wired in series or in parallel, but the principle remains the same.

The lenses are molded into a polyethylene dome that is often white and covers the pyroelectric detector. The dome appears smooth on the outside, but fine patterns of concentric ridges are molded on the inside. These are [fresnel lenses](#), which are much cheaper, smaller, lighter, and easier to fabricate than conventional optical lenses. A fresnel lens introduces some distortion and aberration, but these defects are unimportant in a PIR.

[Figure 4-5](#) illustrates the principle of a simple fresnel lens. The first section of this figure shows a conventional optical lens with one flat side and one curved side. A distant object emits almost-parallel beams of infrared light that are focused by the lens. Section 2 shows the same lens divided into segments that are stacked with no space between them. They behave in exactly the same way as the original lens. In section 3, each segment has been reduced in width, but because the optical faces still have the same geometry, they will still have the same function, although a small amount of distortion will be added by the reduction in width. This is a fresnel lens, which found an early application in lighthouses, where it greatly reduced the weight of very large glass lenses that focused the beam.

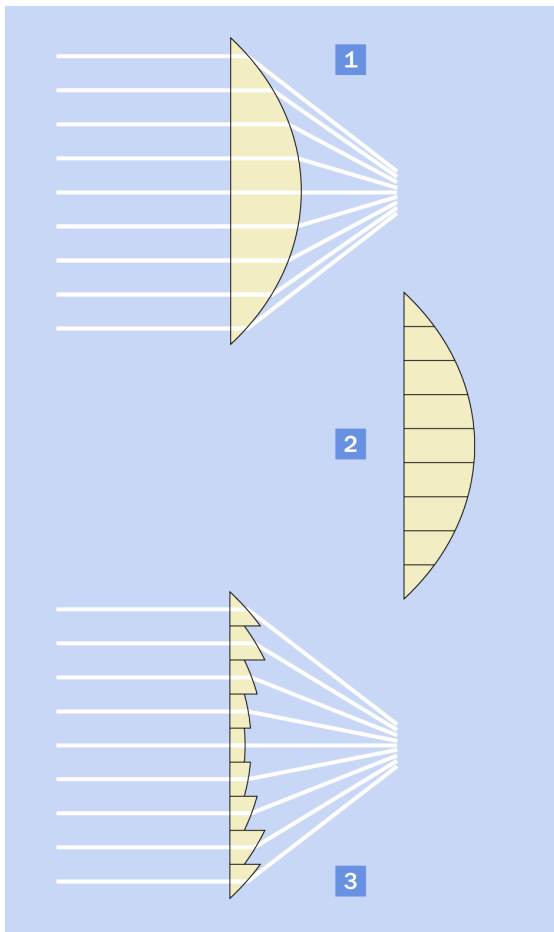


Figure 4-5 Principle of a fresnel lens. See text for details.

The same principle can be applied to a lens in which both surfaces are curved, as shown in [Figure 4-6](#). In practice this will introduce more imperfections in the image, although some compensation is possible by adjusting the exact shape of the lens.

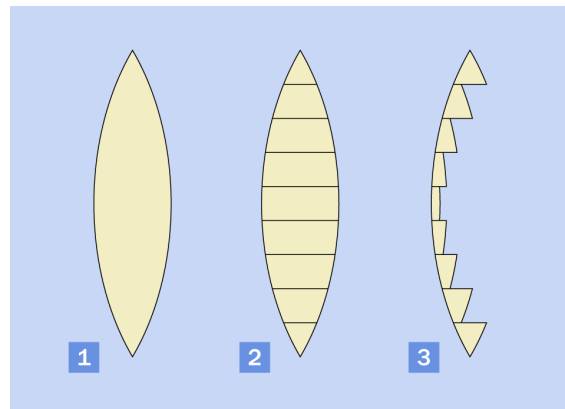


Figure 4-6 The principle of a fresnel lens applied to a conventional lens in which both surfaces are curved.

[Figure 4-7](#) shows three curved fresnel lenses placed edge-to-edge, seen from above. In section 1 of this figure, infrared rays from a distant source are focused by the first lens onto the right-hand element of a pyroelectric sensor. In section 2, the external source has moved laterally, and the rays now focus on the left-hand element. In section 3, the source has moved into the zone covered by the centrally located fresnel lens, which focuses on the right-hand element again. The fluctuating inputs will trigger the sensor.

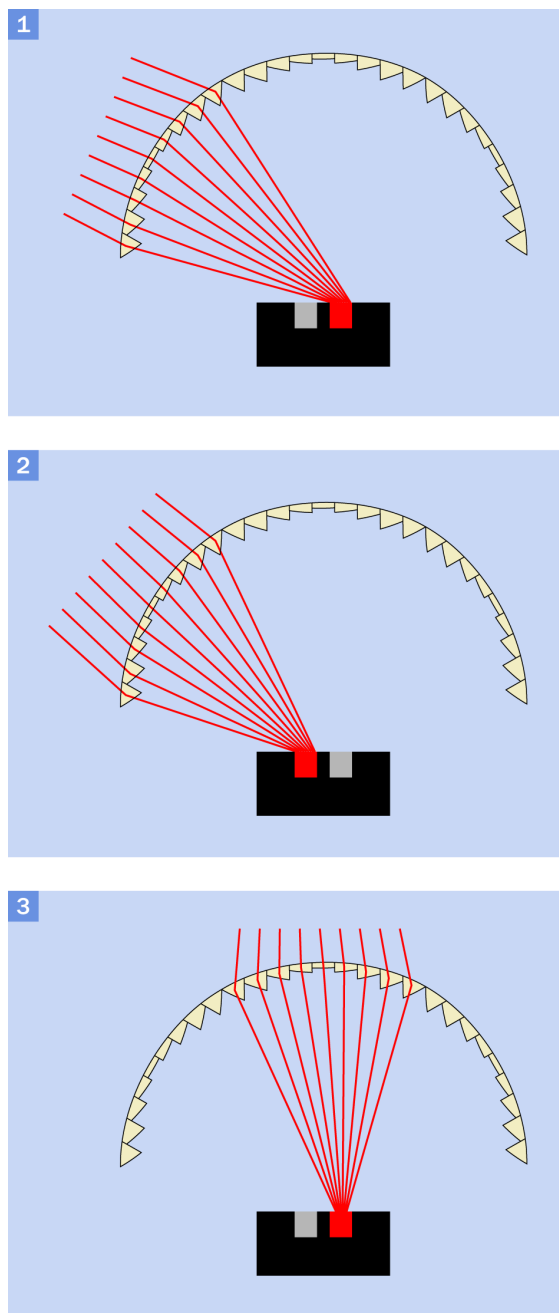


Figure 4-7 Sections 1, 2, and 3 show fresnel lenses focusing an external source of infrared radiation on individual elements of a pyroelectric sensor.

PIRs may combine fresnel lenses in a variety of patterns. [Figure 4-8](#) shows an evenly weighted mosaic that would be suitable for a motion

detector mounted on a ceiling, facing directly downward. [Figure 4-9](#) shows a pattern weighted toward lateral motion, providing less sensitivity for motion above and below the primary band.

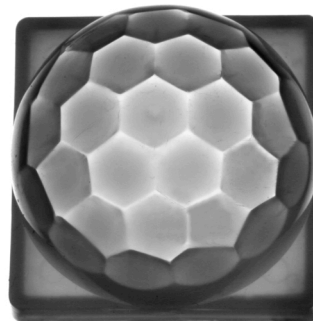


Figure 4-8 An evenly weighted mosaic of fresnel lenses.

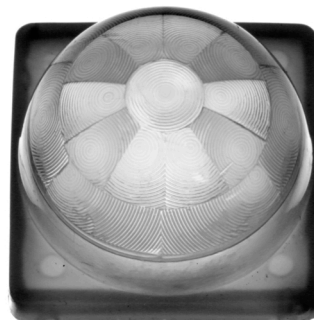


Figure 4-9 A mosaic weighted more toward sensing lateral motion. The grooves of the fresnel lenses are visible.

Variants

PIR sensor modules are available mounted on a small board such as the one shown in [Figure 4-10](#) from Parallax, Inc. The detection range is 5 to 10 meters, selected by a jumper on the board. The three pins visible in the photograph are for power supply (3VDC to 6VDC), ground, and output. The output can source up to 23mA with a 5VDC power supply. Power consumption of the module is only 130μA when it is idle, or 3mA when it is active but has no load.

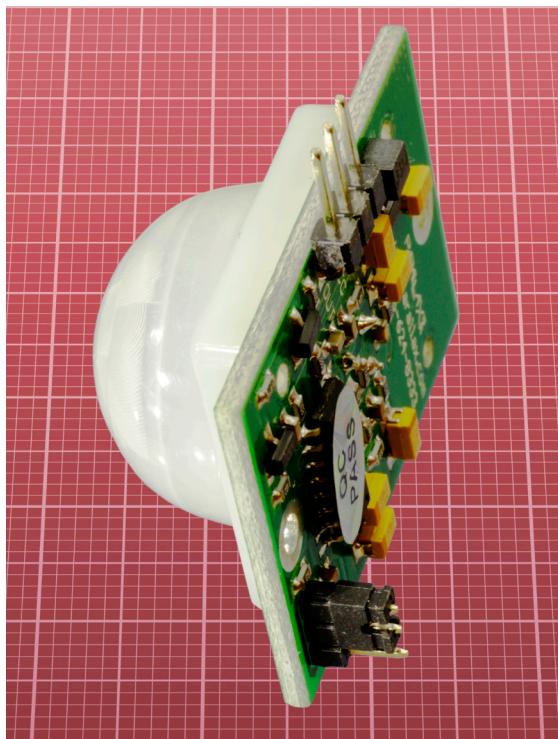


Figure 4-10 A passive infrared detector mounted with basic necessary components on a small board.

A board of this type still requires additional components to set the “on” time for a light or alarm, and to deactivate the PIR during daylight hours.

Various lens patterns are available, sold separately.

A PIR can be bought as a single component containing two elements and FET transistors to amplify the signal. Surface-mount and through-hole versions are available, requiring a typical power supply of 3VDC to 15VDC.

However, a PIR bought as a “bare” component requires significant external circuitry, using comparators or op-amps. Circuit design is non-trivial, entailing practical problems such as op-amps being sensitive to voltage spikes caused by activating a relay that shares the same power supply.

An easier alternative is an all-in-one detector, lens, and control circuit such as the Panasonic AMN31111, which is ready for board mounting. Its small output current of 100 μ A would be capable of activating a solid-state relay. Similar Panasonic PIRs offer a variety of ranges, sensitivities, and supply voltages.

The AMN31111 is in Panasonic’s AMN series. There are many type numbers for combinations of analog or digital output, lens shape, and black or white lens. A selection of lens shapes, derived from the manufacturer’s datasheet, is shown in [Figure 4-11](#).

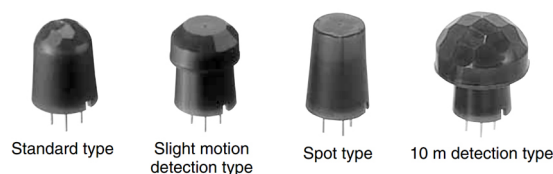


Figure 4-11 Four lenses from the Panasonic AMN series of passive infrared sensors.

What Can Go Wrong

Temperature Sensitivity

In warmer weather, objects in the field of view of a PIR will tend to be warmer, and the temperature difference between them and human skin will diminish. This can degrade the performance of a PIR.

Detector Window Vulnerability

The silicon window on a detector is vulnerable to dirt or grease. Avoid touching the component if it is not protected by lenses.

Moisture Vulnerability

Water absorbs far-infrared light. Consequently, condensation on the lens or detector can degrade performance, and a PIR may not function in heavy rain or snow.

proximity sensor



Proximity sensors that use infrared, ultrasound, and capacitance are described here. This Encyclopedia does not include proximity sensors that use magnetism, inductance, or other methods of determining distance.

Sometimes proximity sensors are referred to as *distance sensors*. An ultrasonic proximity sensor may be described as a *range finder* or a *ranger*.

High-end ultrasound proximity sensors sold as large, sealed modules with cabling can monitor the status of industrial processes. These commercial units are outside the scope of the Encyclopedia.

A sensor that detects whether an object is present, but does not measure the distance to it, is considered to be an **object presence** sensor and has its own separate entry. See Chapter 3. Many devices that are sold as proximity sensors or distance sensors actually do not provide meaningful distance data, and therefore in this Encyclopedia they are included with presence sensors.

Phototransistors and **photodiodes** may be used as sensing elements in proximity sensors. These components have their own entries as light sensors. See Chapter 22 and Chapter 21.

OTHER RELATED COMPONENTS

- **object presence** sensor (see [Chapter 3](#))
- **passive infrared** sensor (see [Chapter 4](#))

What It Does

A proximity sensor measures the distance between itself and a physical object that is often described as the *target*. The output from the sensor may be analog (voltage), serial data, or pulse-width modulation. It may be transmitted via a serial protocol such as SPI, TTL, or I2C, and may be stored as digital data in a register that is accessed by a microcontroller, using I2C. For additional details about protocol, see [Appendix A](#).

Schematic Symbols

Either of the schematic symbols in [Figure 5-1](#) may represent a proximity sensor, but are not used consistently. The sensor may also be represented as a rectangle containing text describing its function.

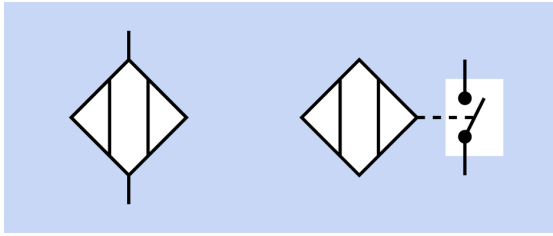


Figure 5-1 Two ways to represent a proximity sensor in a schematic.

Applications

In robotics, a proximity sensor can prevent a collision between a robot and an object or barrier in front of it. Some sophisticated proximity sensors can provide sufficient data for software to map the entire environment, but are outside the scope of this entry.

Proximity sensors can also be used in alarm systems, or for liquid-level sensing in storage tanks, or in automobiles to sound a warning if the driver is backing into an obstacle (although these devices are being augmented with rear-view video monitors).

In handheld devices, proximity sensors are used to sense the presence of the user's hand or face—for example, to shut down the display when a person raises a phone to talk into it.

Variants

This entry is subdivided to describe proximity sensors that use ultrasound, infrared light, and capacitance.

Ultrasound

An ultrasonic proximity sensor functions by emitting a short burst of sound and then listening for echoes from objects in front of it.

The sound is created by a **piezoelectric transducer** (see Volume 2) at a frequency between 30kHz and 50kHz, well above the range that can be detected by the human ear. The transducer may do dual duty as a microphone, sending and receiving sound on an alternating

basis; or a second transducer, serving as a microphone, may be mounted beside the emitter on a small circuit board. The economically priced HC-SR04 is an ultrasound proximity detector popular in the robotics community, working reliably in the range of 2cm to 5m.

A board on which the sensor is mounted may include a microcontroller to measure the delay between propagation of a pulse and reception of the echo. Distance to the reflecting object is then calculated using the speed of sound in air at sea level, which is approximately 340 meters per second.

Infrared

An infrared proximity sensor requires a beam of infrared light from an LED that may be incorporated in the sensing module or mounted separately. Light reflects from the target and is detected by a **phototransistor** or **photo-diode**. From the angle of the reflected light, onboard electronics can calculate the distance to the target by a process known as *triangulation*. See Figure 5-2. (This diagram is simplified. An actual sensor may use a linear array of photodiodes to assess the angle of the returning light beam.)

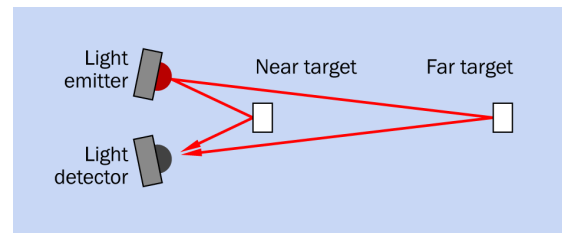


Figure 5-2 An infrared proximity sensor determines the distance to an object by assessing the angle of reflected light.

To reduce the risk of false positives, light from the LED contains only a narrow range of infrared wavelengths. Also, it is modulated at a frequency recognized by sensing circuitry in the module.

Relative Advantages

Ultrasound Devices

- Generally more suitable for detecting objects that are more than 1 meter away.
- Unaffected by direct sunlight, fluorescent tubes, and other light sources that can interfere with infrared devices.
- More accurate, often capable of placing objects within 5mm.
- Able to measure the distance to liquids and transparent objects, which may not be easily detected with infrared.

Infrared Devices

- Physically smaller—especially surface-mount versions.
- Able to measure the distance to soft objects, which may not be easily detected by ultrasound.
- More appropriate for targets that are closer than 10mm.
- More affordable.

Ultrasonic Examples

The proximity sensor in [Figure 5-3](#) is a low-priced model, popular in the robotics community. Manufactured by MaxSonar, it uses a single element to send and receive. The manufacturer claims that it can detect a 6mm (quarter-inch) dowel 1.8 meters directly in front of it, and a 9cm dowel at 3.3 meters. This performance is beyond the capability of almost all infrared sensors.

MaxSonar offers a variety of sensors that appear physically very similar but have different range capabilities. Each has three outputs that can be accessed simultaneously: serial data at 9,600bps, analog voltage, and pulse-width modulation.



Figure 5-3 The MaxSonar MB1003 can detect large solid objects as far away as 5 meters. Photo derived from an image by Adafruit.

The serial output uses the RS232 protocol and consists of the letter R followed by four ASCII-coded numerals representing the measured range in millimeters. Thus, R1000 would indicate an object at 1 meter distance.

The analog voltage ranges linearly from 293mV when sensing an object at 300mm to 4,885mV for an object at 5,000mm.

The pulse-width output sends pulses ranging from 300μs indicating an object 300mm away to 5,000μs indicating an object 5 meters away.

The unit incorporates a temperature sensor that compensates for the lower density of air when its temperature increases. Weatherproofed versions are available. The power supply is 5VDC and must be smoothed.

Imports

Some international sources offer minimal paired ultrasound components at a very low price, such as the HC-SR04 from Cytron Technologies in Malaysia. See [Figure 5-4](#).

Sound dispersion of the transducer is claimed to be plus-or-minus 15 degrees, and the claimed range is up to 4 meters. The module requires a triggering input pulse that must last at least 10μs. This prompts it to emit eight rapid

ultrasound signals at 40kHz. The module measures the response time and applies a high state to its echo pin for a duration that is proportional with the distance measured. An external microcontroller must time the duration, then divide by a factor of 58 to obtain distance in centimeters.

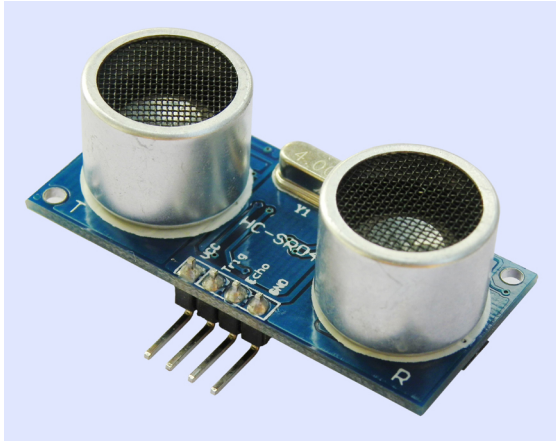


Figure 5-4 The HC-SR04 is a very low-cost import that can provide acceptable performance when used with an external microcontroller.

Many online sources offer simple code libraries for Arduino or PICAXE microcontrollers for use in conjunction with the HC-SR04.

Individual Elements

Individual ultrasonic components such as those shown in [Figure 5-5](#) are available from many vendors. The user must add circuitry to generate a high frequency for a short duration, amplify the microphone signal, measure the time difference, and calculate distance.

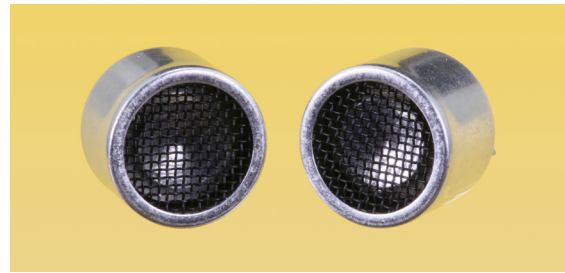


Figure 5-5 These components, listed as the “40TR12B-R ultrasonic sensor kit” by the online supplier Jameco, could form the basis of a DIY ultrasonic proximity sensor.

Infrared Examples

Sharp manufactures four infrared proximity sensors that are widely regarded as accurate and easy to use. They are popular in the robotics community. Their part numbers and ranges are:

- GP2Y0A51SK0F (20mm to 150mm)
- GP2Y0A21YK0F (10cm to 80cm)
- GP2Y0A02YK0F (20cm to 150cm)
- GP2Y0A60SZLF (10cm to 150cm).

The GP2Y0A60SZLF is the most recent product, with the most impressive specification. The GP2Y0A21YK0F is shown in [Figure 5-6](#).

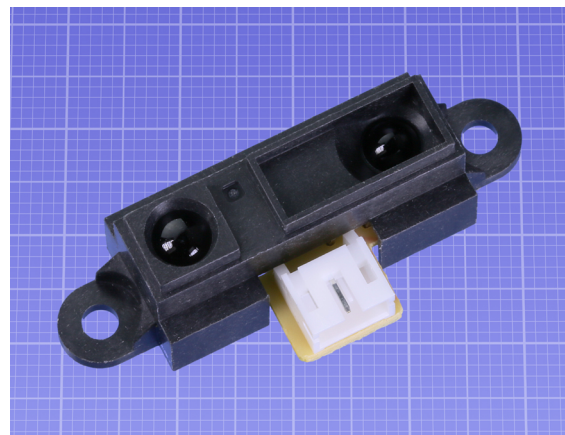


Figure 5-6 The Sharp GP2Y0A21YK0F infrared proximity sensor. The background grid is in millimeters.

Sharp describes these sensors as having an *analog output*. The voltage on an output pin varies in inverse proportion with the distance being measured. The relationship is illustrated by the graph in [Figure 5-7](#), derived from the manufacturer's datasheet for model GP2Y0A02YK0F.

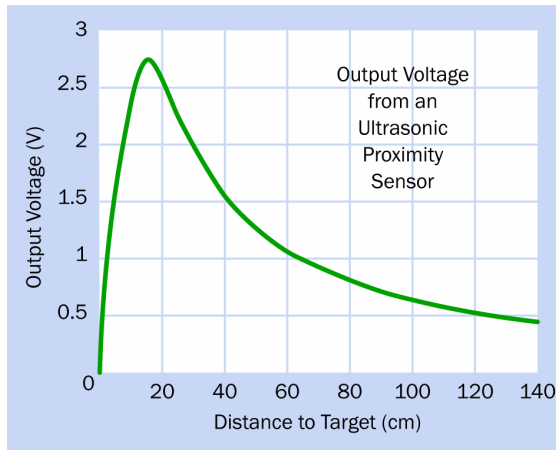


Figure 5-7 Relationship between output voltage and target distance for a Sharp infrared proximity sensor, derived from the manufacturer's datasheet.

The Sharp sensors can work with 5VDC. They consume around 30mA, except for the GP2Y0A60SZLF, which uses less current. Because the infrared LED functions in bursts, the manufacturer recommends protecting other components sharing the power supply by placing a 10μF capacitor across the sensor supply pins.

Trends in Infrared Proximity Sensing

Like many other types of sensors, proximity sensors have been affected by the huge market for handheld devices.

Handheld applications have had four consequences:

Miniaturization

Infrared proximity sensors are now commonly found in surface-mount chips measuring 5mm × 3mm or smaller.

Onboard processing

The status of a photodiode can be processed by a microcontroller on the same chip, to determine what the sensor is really “seeing.” Input from an included ambient light sensor is factored into the evaluation.

Cost reduction

While chip-based proximity sensors have become increasingly sophisticated, their unit cost has plummeted, so that they are now actually much cheaper than simpler devices such as the Sharp analog sensors described above.

Complexity

Modern sensors must be programmed with a complex variety of instructions, and their coded output must be interpreted with a separate microcontroller running its own program. Whether this compensates for the low price and added functionality is something for the individual developer or experimenter to decide.

The Silicon Labs Si1145/46/47 series are chips with the kind of sophisticated capabilities required for handheld devices. An external microcontroller communicates with the sensor via an I2C connection, and can instruct it to adjust its distance range (from 1cm to more than 50cm), its analog-to-digital conversion sensitivity, and its current-sinking capability for up to 3 external LEDs. The chip incorporates ultraviolet sensing and ambient-light sensing capability. Its I2C connection can run at up to 3.4Mbps. Because its light output is pulsed for only 25.6μs every 800ms at 180mA, the average power consumption is only 9μA, assuming a supply voltage of 3.3VDC.

For additional details about protocols such as I2C, see [Appendix A](#).

In addition to its use in handheld devices, the manufacturer suggests applications including heart-rate monitoring, pulse oximetry, and display backlighting control. These applications

require only a subset of the features built into the sensor, but its price is so low, it may be cost-effective even when many of its capabilities are unused.

Proximity sensors with comparable specifications are available from many manufacturers. Examples are the Vishay VCNL4040 and the Avago HSDL-9100. [Figure 5-8](#) shows the Silicon Labs SI1145 on the left, and the Avago HSDL-9100 on the right.

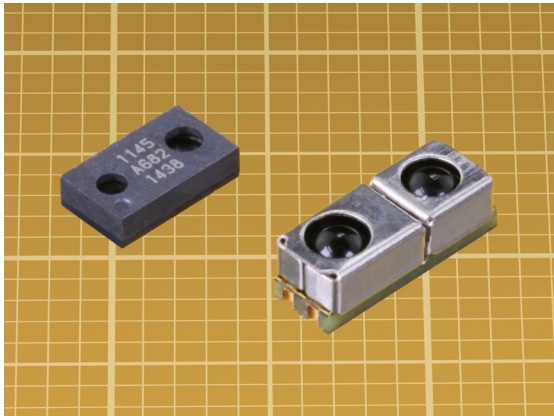


Figure 5-8 Two sophisticated surface-mount proximity sensors with digital output. Left: Silicon Labs SI1145. Right: Avago HSDL-9100. The background grid is in millimeters.

Several guidelines must be observed when using this kind of sensor. First and most obviously, if an external LED is used, it must have a peak wavelength compatible with the photodiode in the sensor. The LED must be placed as near as possible to the photodiode, as reducing the separation increases the sensitivity; but crosstalk between the photodiode and LED must be minimized, usually by placing a thin barrier between them that is barely taller than the higher component.

If the light emitter and/or light sensor are protected behind a transparent glass or plastic panel, it must have minimal resistance to infrared wavelengths, and its thickness must be chosen in compliance with guidelines in the sensor datasheet. To prevent light from reach-

ing the sensor by reflection from the rear of the panel, a thin, opaque tube can be installed between the LED and the panel.

A sensor of this type may be configurable with a “detection scheme,” meaning high and low sensing threshold levels appropriate to the object that is likely to be detected. Determining this may be a process of trial and error.

Capacitive Displacement Sensor

This is also known as a *capacitive linear displacement sensor*. It should not be confused with a capacitive **single touch** sensor, a human-input device that has its own entry. See [Chapter 13](#).

A **capacitive displacement** sensor measures the distance between itself and a target that must be electrically conductive. Unlike optical or ultrasound position sensors, no additional source of light, sound, or other radiation is required. Unlike a magnetic position sensor, it does not require a separate permanent magnet. It simply measures its electrical capacitance with the target.

High-precision capacitive displacement sensors are used mainly for industrial process management. Lower-precision variants are much less expensive and can be used as **object presence** sensors to determine if an object is anywhere within a specified range.

A typical maximum range would be 10mm. For larger distances, optical and ultrasound sensors are more appropriate.

A selection of cylindrical sensor probes is shown in [Figure 5-9](#).

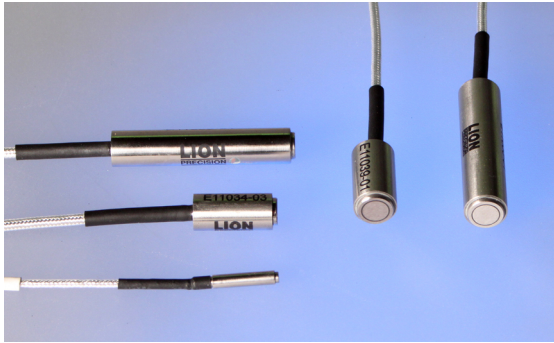


Figure 5-9 Some high-precision capacitive displacement sensors from Lion, which builds them into cylindrical probes ranging from 3mm to 18mm in diameter.

Applications

The high-precision variant of this type of sensor is commonly used during the production of small devices, such as disk drives. It can also measure vibration of a rotating metal part, such as a motor shaft, and may maintain automatic focus of a microscope.

Lower-precision variants can be used for applications such as counting objects on a conveyor.

When used to measure materials thickness, the sensor finds applications in checking automobile brake rotor fabrication and the thickness of silicon wafers.

How It Works

When two plates of electrically conductive material are placed opposite each other, they possess capacitance. This means that they have an electrical storage capacity enabled by accumulation of opposite charges on each plate.

The capacity is directly proportional to the plate area, inversely proportional to the distance between the plates, and is affected by the medium separating the plates, known as the *dielectric*.

If the plate area and dielectric remain constant, the distance between the plates will be the only factor affecting the capacitance. Therefore,

the distance between the plates can be calculated by measuring the capacitance.

Measurement can be performed by evaluating the *displacement current* that passes through the dielectric from one plate to another when a pulse of voltage is applied. (Hence the term “capacitive displacement” in the name of this type of sensor.)

A detailed explanation of displacement current is included in the second edition of *Make: Electronics*.

The sensor itself functions as one plate of the capacitor while the target functions as the opposite plate. Alternating current is applied as a rapid series of pulses, and the current that passes between the plates is proportional to the distance between them.

Ideally, the target should be grounded to the current source. However, since AC is being used, capacitive coupling of the target to the current source is also possible, so long as the additional capacitor has a value of 0.1μF or higher.

Sources of Error

To obtain a meaningful measurement, the electric field from the sensor is focused on the target. Some dispersion will still occur, and the combination of sensor and target must be compatible. A target typically should be flat and should have a larger surface area than the sensor.

Humidity can affect sensor performance, as it changes the value of the dielectric. Temperature can affect performance, partly because it causes small dimensional variations in the sensor and the target.

The surface of the sensor and the surface of the target should be precisely parallel, because the spot where the field hits the target will be elongated if the target is tilted relative to the sensor. Elongation increases the capacitive area and affects the accuracy of measurement.

This type of sensor can also measure the thickness of a nonconductive material, if the material is a thin sheet that can be interposed between two sensors. In this mode the material functions as a dielectric, and its thickness will affect the AC current passing through it.

While lower-precision capacitive displacement sensors are relatively unusual, they can potentially serve as a relatively affordable and simple **object presence** sensor, so long as the target is conductive and will not be damaged by passing a small alternating current at a relatively low voltage.

Values

A high-precision capacitive displacement sensor can measure distances usually ranging from 0.25mm to 10mm with an accuracy that can be better than 0.05mm. High voltages are not required, with a supply of plus-or-minus 15V being common.

The sensing element, often referred to as a *probe*, is usually plugged into a custom-designed control unit that converts the capacitance measurement into a variable output voltage. The performance is then expressed in millimeters per volt. Thus if the voltage varies by 5V over a distance of 1mm, the sensor is rated as providing 0.2mm per volt.

What Can Go Wrong with Optical and Ultrasound Proximity Sensors

Object Too Close

Because both types of proximity sensors (ultrasonic and infrared) may include emitters that are angled toward the distance range for which they are designed, they may fail to “see” an object that is closer. Consequently the sensor will provide no response, or may detect a different object that is further away. In either case, if

the sensor is being used on a moving device, the device may collide with the undetected near object.

Multiple Signals

If two or more sensors and emitters are used concurrently, their combined signals can interfere with each other and create inaccurate readings.

Inappropriate Surfaces

Ultrasonic proximity sensors are intended to identify a single object that is closer to the sensor than other objects, within a narrow beam dispersion angle. Multiple objects, complex surfaces, soft surfaces such as clothing or furnishings, or an unusual configuration of interior walls can create inaccurate readings.

Infrared sensors may be unable to “see” liquids or transparent objects, and may give different assessments of distance depending on the properties of a surface. Human skin, for example, is a poor reflector, as it absorbs some infrared radiation.

Environmental Factors

An ultrasonic transducer uses a very small moving diaphragm to generate sound. Like any system containing moving parts, it will be vulnerable to moisture or excessive humidity, and may need to be protected.

After a device is built and tested indoors, in a controlled environment, it is likely to behave differently if it is moved outside where the temperature is significantly higher or lower.

Deterioration of LEDs

As noted in the section on presence sensors (see [Chapter 3](#)), LEDs tend to suffer from gradual reduction in light output over time. The performance of an infrared proximity sensor may deteriorate over a period of years, depending on how intensively it is used.

linear position sensor



A **linear position** sensor may also be described as a *linear displacement sensor* or a *linear position transducer*. It is sometimes categorized as a type of **proximity** sensor, but in this Encyclopedia a proximity sensor emits a signal and receives an echo to measure the distance from itself to an object. By comparison, a linear position sensor measures the position of a sliding object within a stationary enclosure.

An **object presence** sensor can be considered as a form of linear position sensor, but it only responds to the presence of an object, without measuring its position.

OTHER RELATED COMPONENTS

- **proximity** sensor (see [Chapter 5](#))
- **object presence** sensor (see [Chapter 3](#))
- **rotary position** sensor (see [Chapter 7](#))

What It Does

Control of a mechanical device may require accurate and timely information about the position of a movable part in the device. A **linear position** sensor can be used for this purpose.

Three attributes are likely to be of interest:

- Position
- Direction of motion
- Speed of motion

Typically a linear position sensor measures only the first attribute. Additional electronics may calculate the second and third attributes by taking multiple position readings. Thus a [speed sensor](#) is very likely to be built around a position sensor, and therefore this Encyclopedia does not have a separate entry for speed sensors.

Applications

Robotic arm positioning, wing flap and rudder positions on an aircraft, computer-controlled machine tools, 3D printers, and automobile seat position sensors are some of the many applications.

Schematic Symbol

In a schematic, a linear position sensor may be represented by symbols for the sensing elements that are inside it (potentiometer, LED, phototransistor, or others).

How It Works

Linear potentiometers, *magnetic linear encoders*, *optical linear encoders*, and *linear variable differential transformers (LVDTs)* are described here. Other options are available, but they tend to be more specialized, and are not included.

Linear Potentiometer

For a full description of **potentiometers**, see Volume 1.

A linear potentiometer, often referred to as a *slider potentiometer*, contains an electrical resistance in the form of a straight section of *track*. The track may be a strip of *resistive polymer* or (less often) may consist of an insulator with a coil of nichrome wire wrapped around it.

For sensing purposes, the potentiometer is wired as a voltage divider, and a fixed potential is applied across its full length, as shown in [Figure 6-1](#). A *wiper* slides along the track, sensing a voltage that varies linearly with the wiper's position. Output from the wiper can be used directly to control an analog indicator such as a meter, or can be processed by an analog-to-digital converter.

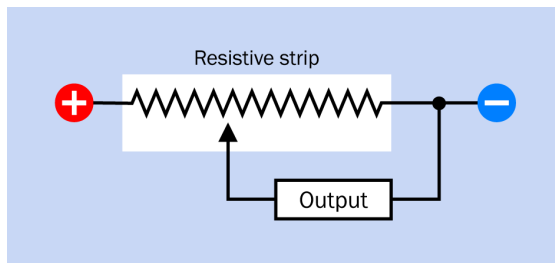


Figure 6-1 A linear potentiometer can consist of a fixed track with a known voltage applied at each end, and a movable wiper.

For audio applications, a slider potentiometer may have resistance that varies logarithmically with the position of the wiper. However, this type of component is not generally used as a position sensor.

For sensing purposes, the potentiometer is usually protected by a long, narrow box or tube through which a rod slides on sealed bearings. An example is shown in [Figure 6-2](#).



Figure 6-2 A linear potentiometer contained in a tube with sealed bearings.

Small linear potentiometers for position sensing are available from companies such as Bourns. The one shown in [Figure 6-3](#) is about 20mm long, and the rod that slides through it has a travel of approximately 10mm. The component is available with resistance values ranging from 1K to 50K, and its power rating is 1/8 watt. The manufacturer claims a life expectancy of 500,000 cycles.

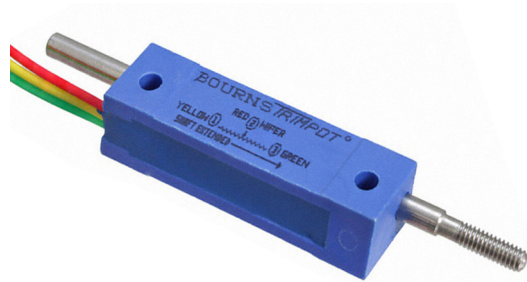


Figure 6-3 A miniature linear potentiometer. The body is about 20mm long.

A linear potentiometer is simple, inexpensive, compact, and requires few additional components. The track contains a lubricant, but inevitably some wear occurs as a result of motion of the wiper. Life expectancy will be reduced by vibration or by contamination with dirt or moisture.

A linear potentiometer may rarely be described as a *linear potentiometric sensor*.

Magnetic Linear Encoders

A ferrous rod or strip can be magnetized with alternating north and south poles. When it slides past a single bipolar Hall-effect sensor, it generates a pulse train from the sensor that can be interpreted to provide positional information. The principle of the component is shown in [Figure 6-4](#). (Magnetic rotary encoders also exist; see “[Rotary Encoders](#)”).

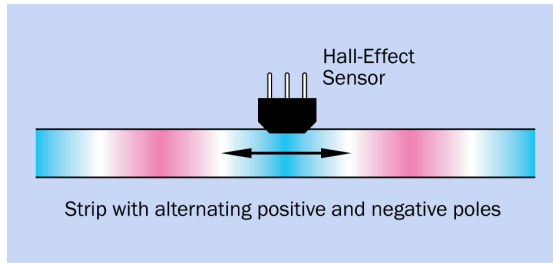


Figure 6-4 When a strip magnetized with alternating north and south poles slides past a sensor, the pulse train from the sensor can be decoded to indicate the relative position of the strip.

The sensing element may be described as a [read head](#). If two are used, with a spacing equal to half the interval between the north and south poles on the strip, the phase difference between the pulse trains from the sensors will indicate the direction in which the magnetized strip is moving. This is shown in [Figure 6-5](#).

The combination of pulse trains is known as [quadrature](#), because there are four possible combinations: A and B both high, A and B both low, A high and B low, or A low and B high. The same principle is used in an optical rotary position sensor; see [Figure 7-6](#).

This type of linear position sensor is often described as a [magnetic encoder](#), meaning that the position of the sliding part is encoded in a series of pulses. Relatively high resolution is possible, as the north and south poles on a strip of ferrous material can be as close as 2mm. Optical encoders may use the same principle; see “[Optical Linear Encoders](#)”.

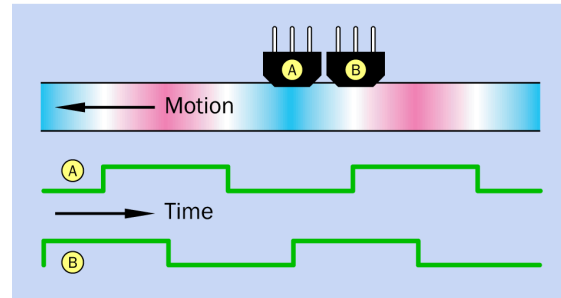
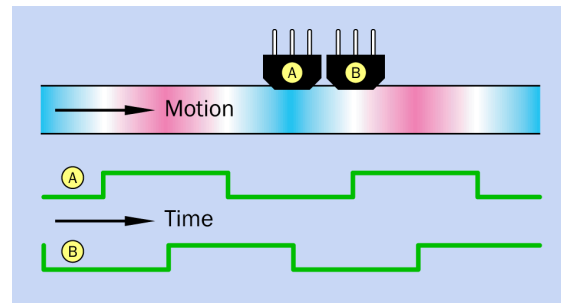


Figure 6-5 Pulse trains from Hall-effect sensors A and B are shown in green. The phase difference between them can be interpreted to show the direction in which the magnetized strip is moving.

The sensor can be built into a module containing an analog-to-digital converter that provides numeric output defining the position of the read head.

In an [absolute](#) magnetic encoder, nonvolatile memory can store the digitized position when the device is switched off. An [incremental](#) magnetic encoder does not store this information, and will require at least one additional [home sensor](#) to detect when the encoder is at either end of its travel. At power-up, an initialization routine moves the magnetized strip until the home sensor is triggered.

For details of Hall-effect sensors see “[Hall-Effect Sensor](#)”.

Optical Linear Encoders

The operation of an optical linear encoder is identical to that of the magnetic linear encoder described immediately above, except that a sliding [optical grating](#) is used in conjunction with a light source and a detection device such

as a **phototransistor** or **photodiode** that functions as the read head. The principle is shown in [Figure 6-6](#). The grating may be described as a *codestrip*.

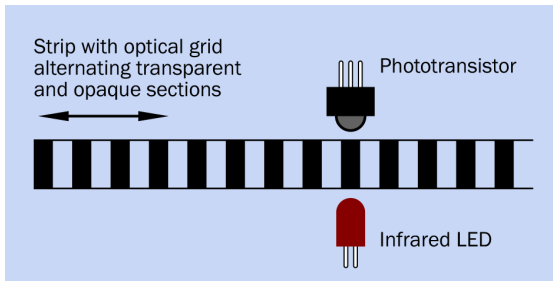


Figure 6-6 An optical linear encoder uses the same general principle as a magnetic linear encoder.

For details of phototransistors, see [Chapter 22](#). For details of photodiodes, see [Chapter 21](#).

An example of a low-cost optical linear encoder is the Avago HEDS-9 series, which consists of a horseshoe-shaped module with an LED in one arm and an array of photodiodes in the opposite arm. When a codestrip passes between the arms, the module emits two pulse trains from internal comparators. The pulse trains are 90 degrees out of phase, and can be interpreted to show which way the codestrip is moving.

These sensors have a body measuring approximately 10mm wide and are designed to read opaque/transparent intervals ranging from 1.5 to 7.87 cycles per millimeter. The output rate can be as high as 20kHz. No pullup resistor is required on the output, as a 2.5K resistor is integrated.

A *codewheel* may be used, in which case the module senses rotation instead of linear motion. This is described in detail in the entry on **rotary position** sensors. See [Chapter 7](#).

Linear Encoder Applications

Optical or magnetic linear encoders are found in some laboratory equipment, machine tools, and industrial robots. The mean time between failures can range from 100,000 to 1,000,000

hours. Optical encoders must be sealed to provide good protection from dust and dirt.

Linear Variable Differential Transformers

This type of sensor, often referred to by the acronym LVDT, tends to be used in industrial environments where great reliability is required under severe conditions. Examples are high-temperature steam valves and nuclear reactor control mechanisms. However, the robust, frictionless design suggests other possible applications, and custom fabrication is a possibility.

[Figure 6-7](#) illustrates the general principle. Three coils are wound sequentially around a (nonmagnetic) stainless-steel tube, enclosed in a second tube also made of stainless steel. The coils act as transformers with the variable voltage ratio determined by the position of a solid iron armature that slides through them.

The center coil is the primary winding, to which AC is applied between 2kHz to 50kHz, depending on the application. (The frequency must be at least ten times the maximum rate of motion of the armature.) The iron armature is attached to a nonmagnetic rod. The voltages on the secondary windings provide the output from the sensor.

While more than one wiring arrangement is possible, the most common schematic is shown in [Figure 6-8](#). The secondary coils are in series, with one of them reversed, so that the phase of the output is inverted as the armature moves from one end of its travel to the other. The phase detector responds to the phase difference by creating a DC output that varies with movement of the armature. All the functions shown in the schematic are available on a single integrated circuit chip.

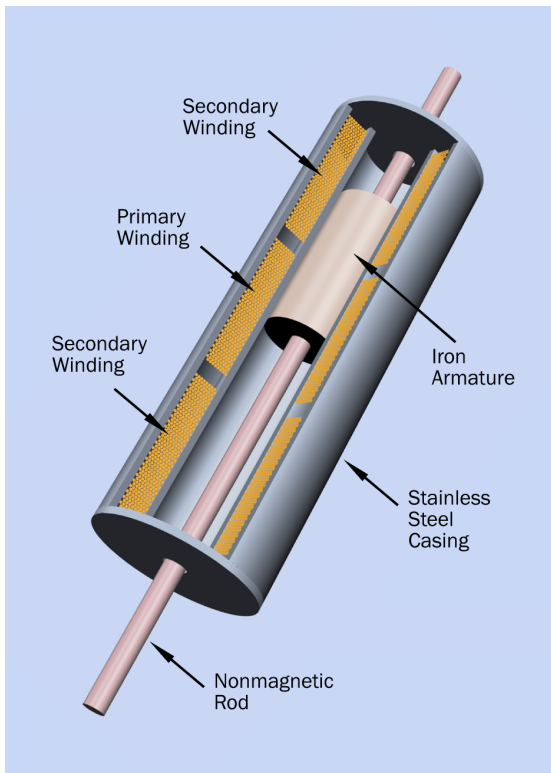


Figure 6-7 Cutaway drawing showing the internal design of a linear variable differential transformer, in which the position of a sliding iron armature determines the voltage induced in the secondary windings.

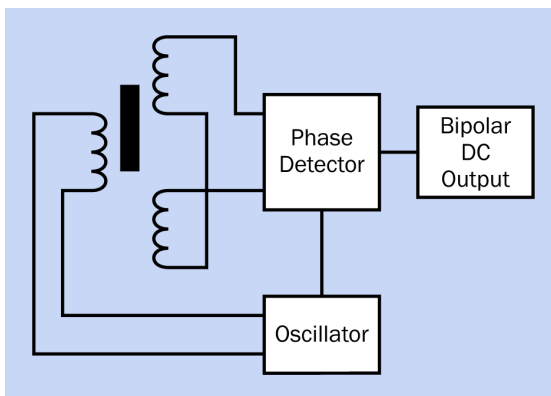


Figure 6-8 Typical schematic for using a linear variable differential transformer.

An example of a linear variable differential transformer is shown in [Figure 6-9](#).

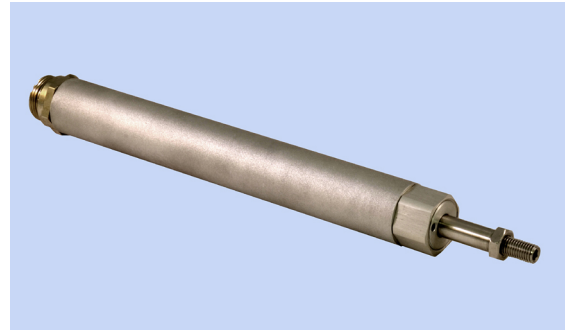


Figure 6-9 External view of a linear variable differential transformer.

What Can Go Wrong

Mechanical Issues

Any sliding mechanism will involve friction, and will be vulnerable to wear and tear, resulting in looseness that will degrade its accuracy. In addition, optical systems are vulnerable to dust and dirt.

LED Longevity

Light output of an LED diminishes over a period of years if the LED is “always on.” This will limit the life of the sensor.

rotary position sensor



Alternative terms for **rotary position** sensor are *rotary sensor*, *rotational position sensor*, *angular position sensor*, and *angle sensor*.

The terms *rotary* and *rotational* are used interchangeably. In this entry, an attempt has been made to use the term that is most common for each specific application. For example, “rotary position sensor” versus “rotational encoder.”

A few sensors are available that specifically measure *rotary speed*, but generally a rotary speed sensor uses information from a rotary position sensor. Therefore, this Encyclopedia does not include a separate entry for rotary speed sensing.

A **rotational encoder** can be used as a rotary position sensor. It is mentioned briefly here but is described in more detail in Volume 1 of this Encyclopedia in the entry describing **switches**.

OTHER RELATED COMPONENTS

- **linear position** sensor (see [Chapter 6](#))

What It Does

Control of a mechanical device may require accurate and timely information about the orientation of a rotating part in the device. A **rotary position** sensor can be used for this purpose.

A sensor may be capable of measuring three attributes:

- Angular orientation
- Direction of rotation
- Speed of rotation

Typically a rotary position sensor measures only the first attribute. Additional electronics are required to calculate the second and third attributes by taking multiple position readings. Thus a *speed sensor* is very likely to be built

around a position sensor, and therefore, this Encyclopedia does not have a separate entry for speed sensors.

Applications

In robotics, a rotary position sensor is commonly used to show the orientation of a pivoting arm or strut. It can also be used as a limit switch on a motor.

Specific applications include solar array positioning, remotely piloted vehicles, guidance and navigation, antenna positioning, and wind turbine pitch control.

Pulses from a rotary position sensor are used to measure speed of rotation in vehicles, industrial processes, and aviation. Small rotary speed sensors are built into devices such as cooling fans and computer hard drives.

Schematic Symbol

In a schematic, a rotary position sensor may be represented by symbols for the sensing elements that are inside it (potentiometer, LED, phototransistor, and others).

Potentiometers

Single-turn and multiturn potentiometers may be used as rotary position sensors. For basic information about **potentiometers**, see the entry describing this component in Volume 1.

Arc-Segment Rotary Potentiometer

An *arc-segment rotary potentiometer* is often referred to simply as a “potentiometer,” as this type is more common than the multiturn type or the linear type. When used as a sensor, it can measure a turn angle that is less than 360 degrees.

This component contains a resistor in the shape of an arc, referred to as the *track*. It may be a strip of *resistive polymer* or (less often) may consist of an insulator with a coil of nichrome wire wrapped around it.

For sensing purposes, the potentiometer is wired as a voltage divider, and a fixed potential is applied along the full length of the track, as shown in Figure 7-1. A *wiper* slides along the track, sensing a voltage that varies linearly with the wiper’s angular position. Output from the wiper can be used directly to control an analog indicator such as a meter, or can be processed by an analog-to-digital converter.

For audio applications, an arc-segment potentiometer may have resistance that varies logarithmically with the position of the wiper. However, this type of component is not generally used as a position sensor.

Low-cost potentiometers were traditionally used as volume or tone controls in stereo systems. When a potentiometer is designed for use as a sensor, it tends to be more ruggedly

built and better protected against dust, dirt, and moisture. Its advantages are that it is simple, inexpensive, compact, and requires few additional components.

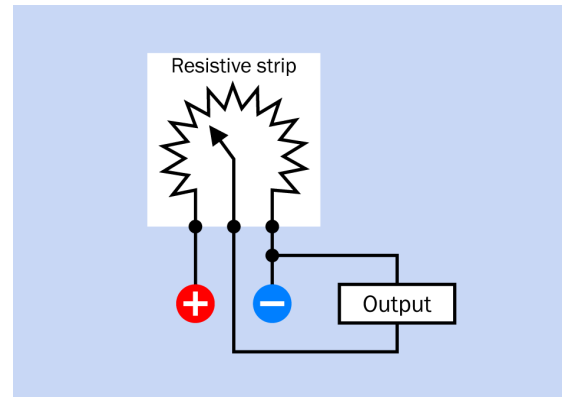


Figure 7-1 An arc-segment potentiometer consists of a fixed track in the shape of an arc with a known voltage applied at each end, and a movable wiper.

The major disadvantage is that although the track contains a lubricant, some wear will gradually result from friction with the wiper. Life expectancy will be further reduced if there is vibration or contamination with dirt or moisture.

End Stops

An arc-segment potentiometer usually has *end stops* to prevent the wiper from running off either end of the track. Typically, these stops limit rotation to around 300 degrees.

A few arc-segment potentiometers allow unrestricted rotation. The Bourns 6639 series is an example, although the wiper still passes through a “dead zone” of 20 degrees between the start and end of its track. An application for this type of potentiometer could be as a direction sensor for a weather vane.

Multiturn Rotary Potentiometer

A spiral track, which resembles a coil spring, enables a rotary potentiometer to make multiple turns. The wiper rotates inside the track and follows its contour. A potentiometer of this type

will still be calibrated in degrees; thus, a 10-turn component may be listed as allowing an *electrical travel* of approximately 3,600 degrees.

The exterior of a multiturn rotary potentiometer is shown in [Figure 7-2](#).



Figure 7-2 A multiturn rotary potentiometer. Each pair of solder tags is attached internally to one end of the internal coiled track. The smaller tag at the far end connects with the wiper.

A simpler type of multiturn potentiometer is intended for use as a trimmer—a small potentiometer that can be mounted on a circuit board to allow adjustment or calibration, often during the manufacturing process. This type of trimmer contains a worm gear that engages with a spur gear internally. The wiper is mounted on the spur gear. It has no applications as a sensor, but is mentioned here to avoid ambiguity, as it is probably the component that is most commonly referred to as a “multiturn potentiometer.”

Magnetic Rotary Position Sensor

Externally, a modern magnetic rotary position sensor may look very much like an arc-segment rotary potentiometer. Internally, a permanent magnet is attached to the base of the shaft, and one or more **Hall-effect** sensors are mounted on a small circuit board immediately below the magnet, in the bottom of the enclosure. A simplified diagram appears in [Figure 7-3](#).

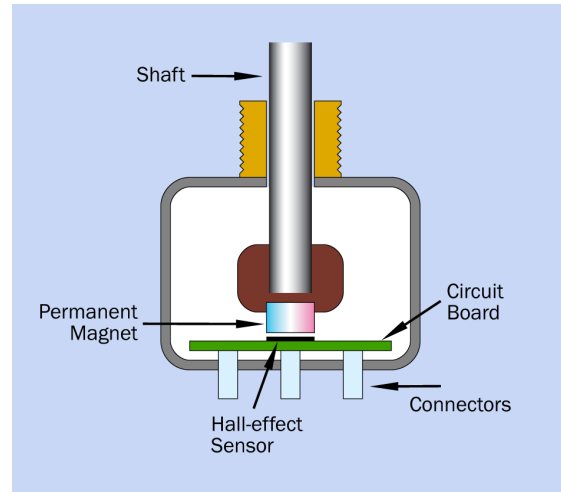


Figure 7-3 Simplified interior view of a magnetic rotary position sensor.

For more information about Hall-effect sensors, see “[Hall-Effect Sensor](#)”.

A magnetic rotary position sensor may be described as a *noncontacting* sensor. Two views of an example are shown in [Figure 7-4](#).

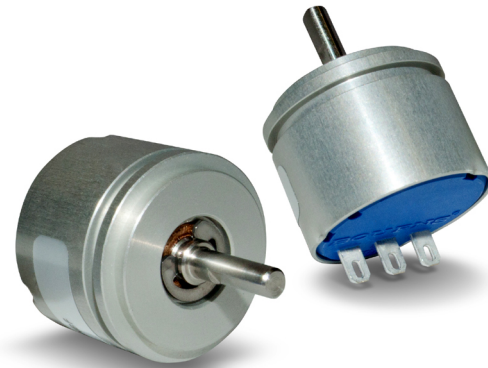


Figure 7-4 Two views of a Bourns AMS22 magnetic rotary position sensor.

The AMS22 sensor shown in [Figure 7-4](#) has an analog output ranging from 0.1VDC to 4.9VDC when powered with 5VDC. A noncontacting sensor of this type will cost three to four times as much as a conventional potentiometer, but has the great advantage of extreme durability, with the manufacturer claiming a life of 50 mil-

lion shaft rotations. A disadvantage is its low-current output, limited to 10mA. The maximum rotation speed of its shaft is 120rpm.

Rotary Position Sensing Chips

Chips of the type used in a magnetic rotary position sensor are available as individual components, many of which have advanced features. For example, the AM8192B angular magnetic encoder by RLS is a 44-pin surface-mount chip containing Hall sensors that detect the orientation of a permanent magnet above or below the chip. Various outputs provide information such as sine or cosine of the turn angle, incremental pulses, and digitized output via an SPI interface.

Rotary Encoders

Many rotating position sensors communicate their angle of rotation with an output that consists of a pulse train or some other coded signal. This type of rotary sensor is known as a *rotary encoder* or *rotational encoder*. (Linear encoders also exist; see “Magnetic Linear Encoders”.)

The simplest version of this component is a *mechanical encoder* containing two small electromechanical switches that are activated, out of phase, by a toothed wheel attached to the rotating shaft. This component is described in detail in Volume 1 of the Encyclopedia, where it is categorized as being a form of **switch**. Its low cost and simplicity has made it popular for rotary controls on car radios and small stereo systems, but its switches have a limited life expectancy and create a “noisy” output that must be *debounced* if connected with a logic chip or microcontroller. Typically the microcontroller will include a pause of up to 50ms in its program code to allow time for the switch contacts to settle (although some manufacturers claim only 5ms).

Confusingly, a mechanical encoder is often identified only as a “rotational encoder,” even though optical and magnetic rotational encod-

ers also exist, as described immediately below. As a general rule, if a component is described simply as a rotational encoder, it probably contains electromechanical switches.

Optical Rotary Encoders

This type of component works on the same principle as an *optical linear encoder* of the type described in “Optical Linear Encoders”. The difference is that a *codewheel* is used instead of a *codestrip*. Typically the codewheel is supplied by the manufacturer of the component that is designed to read it.

A *transmissive* codewheel is shown in Figure 7-5. The distance between the light emitter and the light detector has been exaggerated in this diagram for clarity.

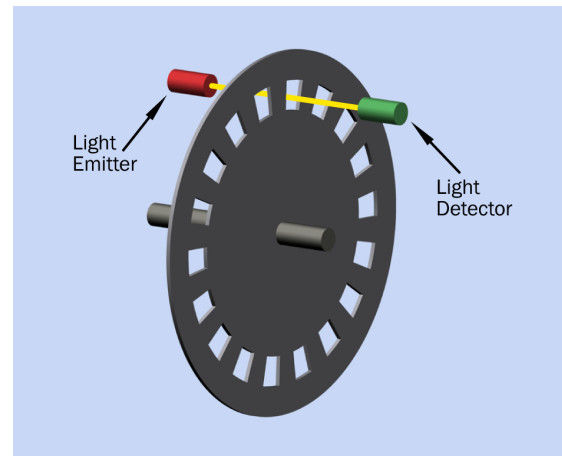


Figure 7-5 A light-transmissive codewheel for use in an optical rotary encoder.

Some optical rotary encoders use a *reflective* codewheel, in which case alternating sections of the wheel are light-absorptive and light-reflective, and the emitter and detector are both on the same side of the wheel.

If only one light emitter and light detector are used, the pulse train from the sensor reveals how many increments the wheel has turned relative to its previous position.

A single sensor cannot indicate the direction of rotation, but if a second emitter-sensor pair is added, 90 degrees out of phase with the first, a microcontroller can assess the phase difference between pulse streams to determine which way the wheel is turning. This principle is illustrated in [Figure 7-6](#) where one transmitter-detector pair is located in position A and another is in position B. The resulting pulse trains, for clockwise and counterclockwise rotation of the wheel, are shown in green. The combination of pulse trains is known as *quadrature*, because there are four possible combinations: A and B both high, A and B both low, A high and B low, or A low and B high.

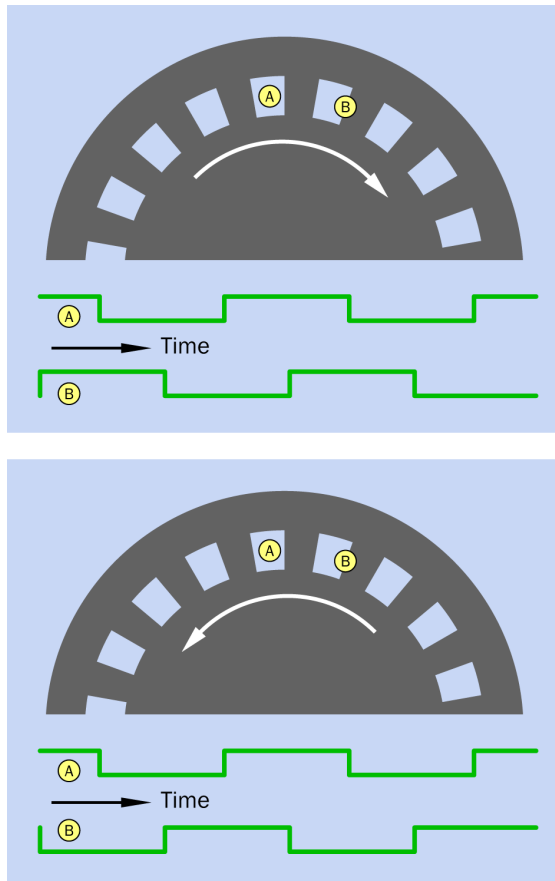


Figure 7-6 With transmitter-detector pairs mounted at positions A and B, the phase difference between their pulse trains can show the direction of rotation of the wheel.

This is the same principle as illustrated for a magnetic linear position sensor in [Figure 6-5](#).

An alternative system of quadrature for an optical rotary encoder uses two separate tracks on the disc, each having an equal number of opaque and transparent sections, but half an interval out of phase.

Any system that reveals the relative motion of the wheel, but cannot determine its absolute angular position, is an *incremental sensor*.

Optical Products

High-end optical rotary encoders may use discs with as many as 600 sequential opaque and transparent segments to provide extremely high resolution. They are outside the scope of this Encyclopedia.

Moderately priced optical encoders are available as shaft-driven assemblies very similar in external appearance to potentiometers. Typically they have four terminals, one pair for power and ground connections and another pair for quadrature output from the two internal sensors, usually identified as A and B in datasheets. Some encoders also contain an on-off switch that is activated by pressing the shaft, in which case two additional terminals will be provided.

Bourns is a leading manufacturer of this type of encoder, an example being the EM14, which is mounted in a box-shaped body measuring 14mm square. It uses a 5VDC power supply and provides pulses of 4VDC minimum with intervals of 0.8VDC maximum. Variants are available with 8 to 64 pulses per revolution. Intended for audio applications, this type of encoder has a maximum rotation speed of 120rpm.

An example of an optical rotary encoder from a German manufacturer is shown in [Figure 7-7](#). Its resolution is 25 pulses per rotation. The rectangular package measures approximately 19mm × 25mm. The power supply can be 3.3VDC or 5VDC.

Optical rotary encoders of this type cost about five times as much as mechanical rotary encoders at the time of writing, but their longevity and their clean output signals make them an attractive alternative, and the price difference may diminish over time.

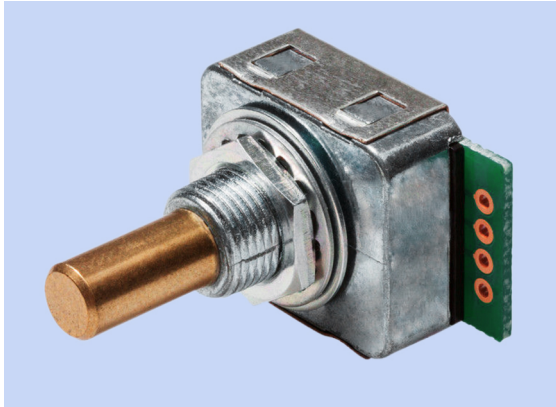


Figure 7-7 A compact incremental optical rotary encoder in the MRB25 series from Megatron Elektronik AG & Co.

Encoders such as the Avago HEDS-9 series are not protectively enclosed, and require assembly with a codewheel supplied by the manufacturer. See [Figure 6-6](#) for additional information.

A more basic optical rotary encoder, shown in [Figure 7-8](#), is sold by Cytron Technologies as an *optical switch* mounted on a small board, with a separate codewheel consisting of a slotted disc. This low-cost kit is intended for use in DIY robotics. More information about optical switches will be found in the entry dealing with **optical presence** sensors. See [Figure 3-3](#).

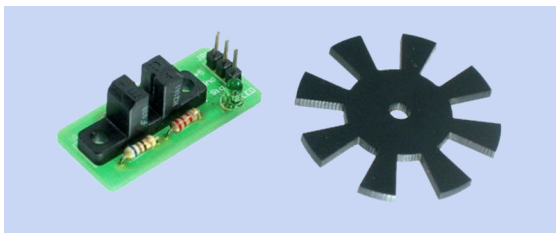


Figure 7-8 A bare-bones optical encoder for DIY robotics.

Computer Mouse Principles

The original design of a computer mouse, with a hard rubber ball, contained two optical rotary encoders oriented at right angles to each other. Each of them used a transmissive codewheel. The rolling ball turned the codewheels as the mouse was moved across a desktop, and electronics in the mouse converted the outputs from the encoders into a pulse train that could be interpreted by a computer. [Figure 7-9](#) shows the primary components.

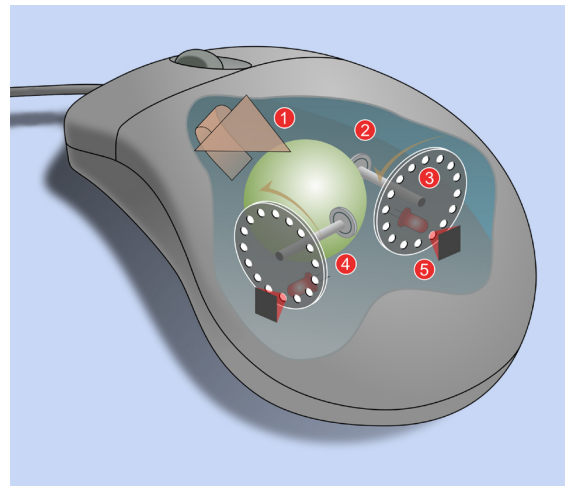


Figure 7-9 From Wikimedia Commons, this rendering by Jeremy Kemp shows (1) rotation of the ball, (2) a roller touching the ball, (3) a transmissive optical codewheel, (4) an infrared LED that shines through a second codewheel, and (5) a sensor detecting the pulses of light.

An *optical mouse* works on a different principle, maintaining a monochrome image of the desk surface on an optical array that functions like a very low-resolution camera sensor. Electronics in the mouse detect displacement of the image as the mouse is moved.

Rotational Speed

Incremental rotary encoders, which only supply relative data, are adequate in many applications, especially speed measurement, where a microcontroller can compare a pulse stream from a sensor with a target frequency, and can also provide feedback to control motor speed

appropriately. This is convenient when a stepper motor is being used, or a DC motor that is controlled with pulse-width modulation. (See the entry on **motors** in Volume 1.)

Toothed wheels or magnetized wheels are commonly used to measure speed of rotation in applications ranging from automobile transmissions to computer disk drives.

While a second sensor can be added to determine the direction of rotation, this still does not provide information about the absolute orientation of a part.

Absolute Position

If associated electronics are equipped with nonvolatile memory, it may be used to store a sensor-wheel position from one session to the next. This may be sufficient in noncritical applications such as volume control in a car radio or stereo system.

Alternatively, an additional single window in an optical wheel can activate a [home sensor](#). When the device is powered up, the wheel is turned until the home sensor is triggered, at which point the orientation of the wheel is known, and subsequent pulses from rotation sensor(s) will add or subtract angular information.

The pulse generated by a home sensor may be described in datasheets as a [reference signal](#) or [index signal](#). In the early days of desktop computers, each 5.25-inch floppy diskette was perforated with an index hole for this purpose.

The Gray Code

For greater reliability in determining absolute position, an optical wheel can be divided into several concentric tracks, each of which contains a different coded sequence and is assigned its own light emitter and light sensor. The sensors are arranged in a radial line to scan the disc as it rotates. Since each detector will provide either a signal or no signal, output from the set of sensors can be combined to create a binary number.

Figure 7-10 shows binary codes from 0000 through 1111, corresponding with decimal numbers 0 through 15, where a white square is equivalent to a numeral 1 and a black square is equivalent to a numeral 0.

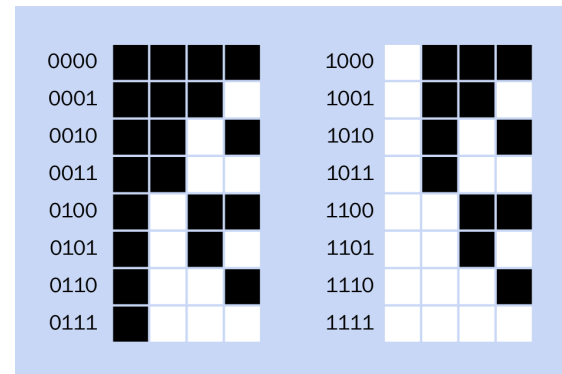


Figure 7-10 Binary codes from 0000 through 1111, using white to represent 1 and black to represent 0.

Figure 7-11 shows this system mapped onto a codewheel. The red circles indicate the locations of four stationary light detectors, which begin by providing a reading of 0000, since they coincide with four opaque areas of the wheel. Now if the wheel makes 1/16th of a full rotation in the direction of the arrow, the detectors will register 0001. If the wheel continues to rotate, the detectors will count in binary up to 1111 before the sequence repeats.

The problem with this design is that small manufacturing inaccuracies and other imperfections will result in some light detectors responding fractionally more quickly than others as the wheel rotates. This will occur in transitions where two or more adjacent segments of the wheel change between transparency and opacity. For instance, where 0011 is followed by 0100, transient values of 0010, 0001, 0111, 0110, or 0101 are possible. Although they will be brief, these values may trigger associated electronics.

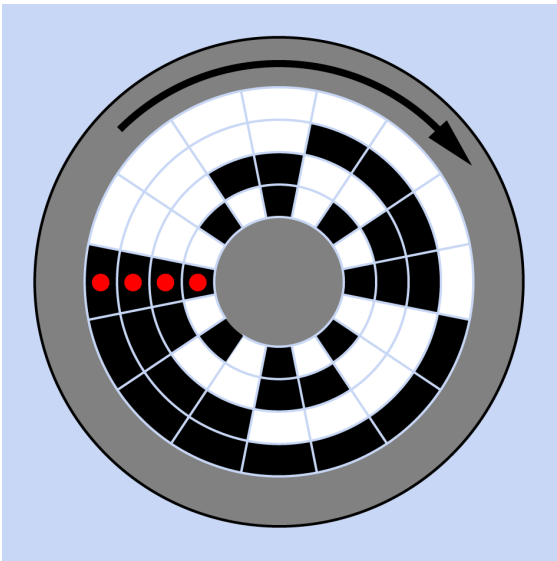


Figure 7-11 The binary sequence mapped onto a code-wheel as areas of opacity and transparency. Red circles indicate light detectors.

To eliminate this problem, a different code sequence can be used in which only one of the four sensors is allowed to make a transition from each value to the next. This is called a *Gray code*, and eliminates the issue of simultaneous transitions. A commonly used Gray code is shown in Figure 7-12.

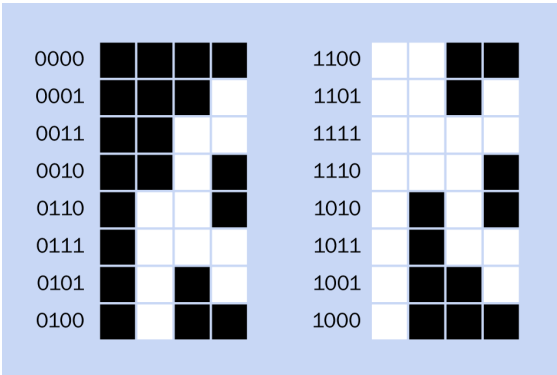


Figure 7-12 A Gray code that allows only one binary digit to change in each transition from one value to the next.

Magnetic Rotary Encoders

If a ferrous wheel is magnetically polarized in multiple domains, its rotation can be detected

by a Hall-effect sensor in much the same way that a wheel divided into transparent and opaque sections can be assessed by light detectors and emitters. This is illustrated in Figure 7-13, where magenta and cyan bands indicate north and south magnetic poles.

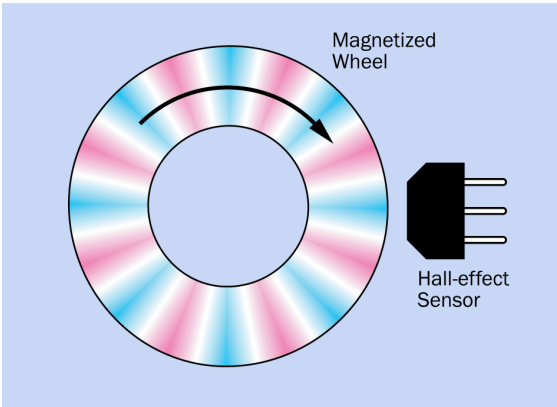


Figure 7-13 A Hall-effect sensor can detect the rotation of a wheel that is divided into multiple north and south magnetic poles.

See “Hall-Effect Sensor” for information about Hall-effect sensors.

An additional Hall sensor can be added, offset from the first, as in Figure 7-14. Once again the phase difference between the pulse trains can be used to determine the direction of rotation.

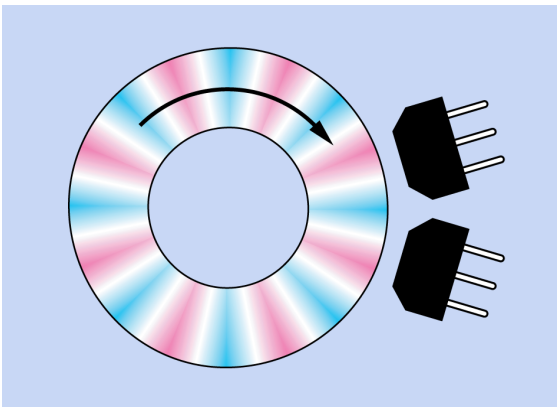


Figure 7-14 Direction of rotation of the wheel can be deduced from the phase difference between the pulse trains from two sensors.

Alternatively, a toothed wheel can be used, as suggested in [Figure 7-15](#).

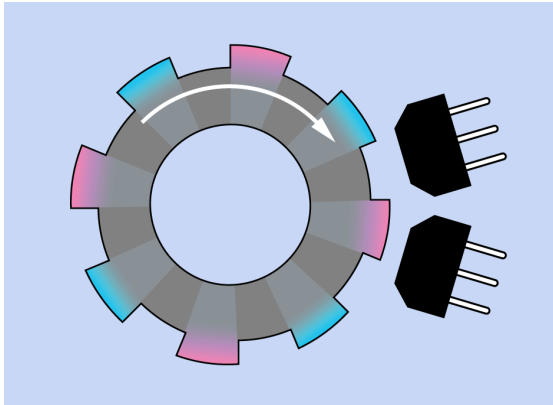


Figure 7-15 A toothed wheel can trigger a Hall sensor if the teeth are magnetized.

Another option is to mount a pair of magnets on a nonmagnetic wheel, with a Hall-effect sensor in the center. If it is a sensor with an analog output, the voltage will fluctuate smoothly between positive and negative, relative to the power supply for the sensor. Alternatively, a bipolar Hall-effect sensor can be used to provide a binary output. The concept is illustrated in [Figure 7-16](#). An advantage of this configuration is that it provides approximate information about the absolute position of the wheel.

The two magnets will provide usable linear sensor outputs over a span of plus-or-minus 30 degrees of rotation, approximately. Additional magnets or sensors can produce a more complex output that would be decoded by a microcontroller.

How to Use It

An optical or rotary encoder is well suited for use with a microcontroller program that can count pulses, compare pulse trains, or interpret a Gray code. The microcontroller then takes appropriate action. For example, if a rotational encoder is used to control the gain of an audio amplifier, the microcontroller determines the direction and angle through which the encoder

turns, and can respond by changing the value of a **digital potentiometer** (see Volume 1).

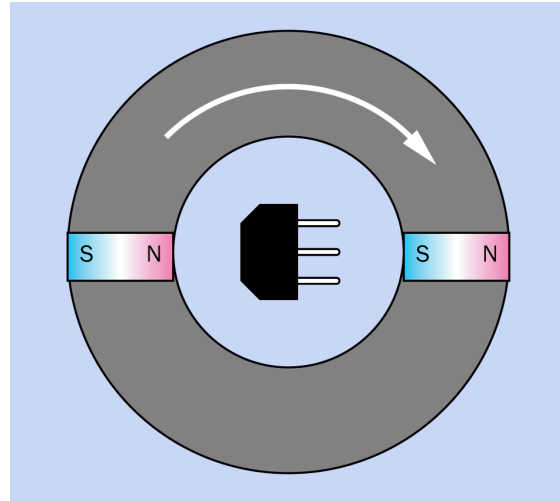


Figure 7-16 A Hall-effect sensor can detect the angle of rotation of a ring on which two magnets are mounted as shown.

Integrated circuit chips are available to convert a sequence of quadrature signals into an up-pulse or a down-pulse, thus eliminating the chore of achieving this with a microcontroller program. The LS7183 by LSI Computer Systems is an example.

What Can Go Wrong

Wiring Errors

If two sensors are used to detect the direction of rotation of a rotary encoder, and outputs from the sensors are accidentally swapped, the component will seem to work normally except that the indicated direction will be inverted.

Coding Errors

Microcontroller code that interprets quadrature signals must be fast enough to keep up with the pulse streams. If a microcontroller is performing other tasks, an interrupt may be necessary for processing the rotational data. This should not be a significant problem when interpreting human input from a knob or dial, but

for a motor-driven encoder, pulse counting in hardware may be a better alternative.

Ambiguous Terminology

Rotational encoders are often referred to simply as “encoders,” because they are more com-

mon than linear encoders. When searching for an encoder, check each datasheet carefully to determine which type you are dealing with.

tilt sensor



A *tilt switch* is defined here as an electromechanical switch, while a *tilt sensor* uses electronics. Both types are included in this entry.

A *tipover switch* is very similar to a tilt switch and uses the same principles. Therefore, it is included in this entry.

Some manufacturers' datasheets refer to tilt sensors as *tip sensors*. Some Asian supply catalogues refer to tilt switches as *breakover switches*.

An **accelerometer** can measure the angle at which it is held relative to the downward force of gravity, but it has additional capabilities. Therefore, it has its own separate entry in this Encyclopedia.

An *inclinometer* measures the incline, or positive slope, from an observation point to the top of an object such as a building or tree. The height of the object can be calculated from the angle. A *clinometer* can additionally measure a decline, or negative slope. These measurement devices are fully featured products as opposed to sensors, and are not included in this Encyclopedia.

OTHER RELATED COMPONENTS

- **accelerometer** (see [Chapter 10](#))
- **vibration** sensor (see [Chapter 11](#))

What It Does

Three principal types of tilt sensor exist.

1. Single axis, single output. The sensor responds to being tilted around one horizontal axis, relative to the downward force of gravity.
2. Dual axis, dual output. The sensor contains two sensing elements at 90 degrees to each other. Each has an output determined by its angle of tilt from vertical around one axis.

3. Dual axis, single output. A single sensor responds to an angle of tilt from vertical around any horizontal axis.

A *tilt switch* is usually of the third type, and is defined here as containing an electromechanical or electronic switch that opens or closes a connection. Most tilt switches are SPST (normally open) or SPST (normally closed). A minority are DPDT.

A *tipover switch* is a type of high-current tilt switch that cuts power to a device such as an electric heater when it is tipped over.

This Encyclopedia defines a *tilt sensor* as being an electronic component, as opposed to an electromechanical component. The distinction is often observed in datasheets, but not always.

Schematic Symbol

No specific schematic symbol is generally used for any variant of a tilt sensor. It can be represented with an annotated switch symbol.

How It Works

Because a tilt switch is a simpler device than a tilt sensor, it will be described first.

The most common type of tilt switch consists of a cylindrical metal or plastic enclosure, often measuring about 5mm by 15mm, containing two spherical steel balls that may be nickel-plated or gold-plated. When the switch is tilted, the balls eventually run downhill, and the lower ball completes an electrical connection between two contacts or between a single contact and the metal enclosure of the switch. The second ball is included to add weight and suppress vibration in the first.

Figure 8-1 shows a switch manufactured by Comus Global, rated for 0.25A at 60VAC or 60VDC, maximum. The body of the switch measures approximately 10mm x 5mm, and the switch is activated when tilting -10 degrees from horizontal. It is deactivated when tilting +10 degrees. A scale drawing of the interior is shown in Figure 8-2.

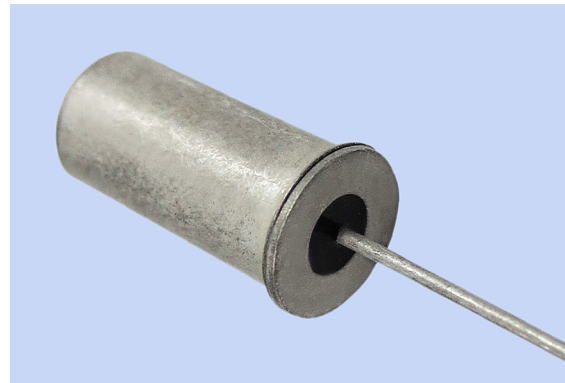


Figure 8-1 The CW1300 tilt switch manufactured by Comus Global.

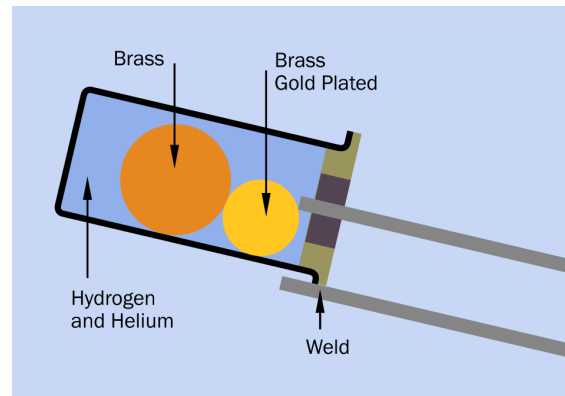


Figure 8-2 Interior of the CW1300 tilt switch, from a scale drawing supplied by the manufacturer. The lower lead is welded to the shell of the sensor. The leads may be inserted in a circuit board.

Figure 8-3 shows three common internal configurations of a generic tilt switch. The top version has axial leads and uses the metal shell of the switch to complete the circuit. The center version has radial leads, with a plastic shell. The bottom version has radial leads, one of which is attached to the metal shell to complete the circuit.

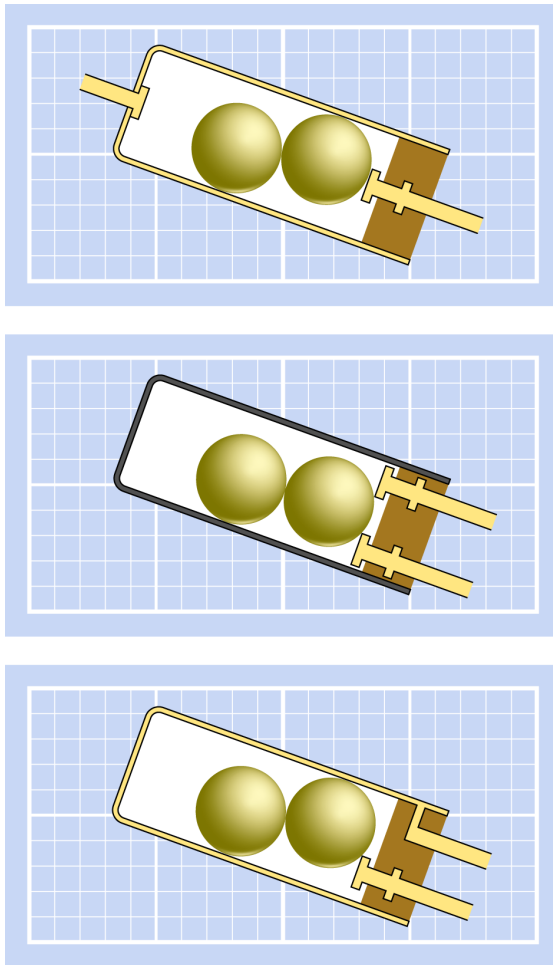


Figure 8-3 Three variants of a generic ball-operated tilt switch. The graph-paper scale is in millimeters.

Figure 8-4 shows the parts of a disassembled tilt switch.

Simplified Version

Where size is not a significant consideration, a tilt switch can be created by attaching a pivoting, weighted arm to a small snap-action switch.

Applications

An old-style (nonelectronic) thermostat may contain a tilt switch attached to the end of a bimetallic strip coiled into a spiral. When the strip bends in response to a drop in temperature, the switch closes its contacts, activating a

relay that starts a heating unit. If the temperature rises, an additional set of contacts in the same relay may activate an air-conditioning unit. In old thermostats, the tilt switch may contain mercury in a glass tube, which should be handled with caution.

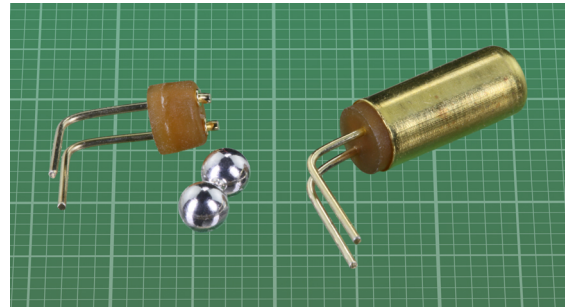


Figure 8-4 At right, a tilt switch. At left, the cap removed, and the two balls that make internal contact. The background grid is in millimeters.

A tilt switch may detect the opening of a door or window in a simple alarm system.

Tilt switches have been used in automobiles to switch on the interior light in the trunk when its lid is opened.

A normally closed tilt switch is often used to stop the inflow of granular material to a bin when it is almost full. This is colloquially known as a [bin switch](#). In industrial applications of this kind, the switch is activated by a long lever that has a ball mounted at the end. The switch assembly is physically large. See [Figure 8-5](#).

A normally open tilt switch may operate a valve or start a pump when the liquid in a tank drops below a certain point. If the switch uses a float to sense the level of the liquid, it is often known as a [float switch](#). This is described in the entry on liquid level sensors. See [Chapter 15](#).

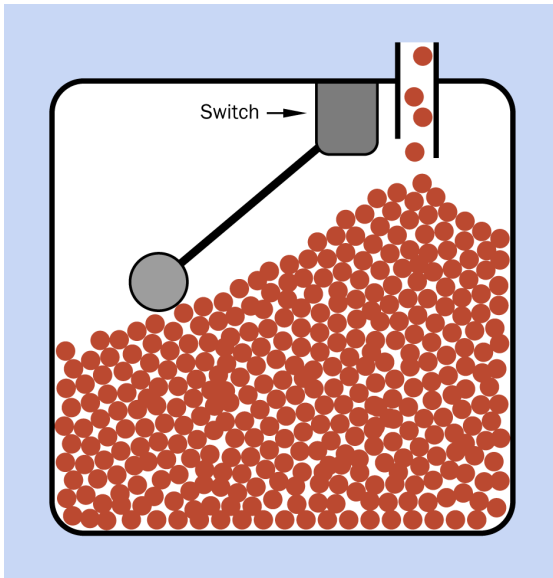


Figure 8-5 The flow of granular material into a bin can be sensed and stopped by a tilt switch of this type, often known as a bin switch.

A *tipover switch* may use the simplified system of a weighted arm that activates a snap-action switch. When used in conjunction with a room heater, the switch must handle substantial current.

A motorcycle may contain a tipover switch to stop the electric fuel pump if the motorcycle falls on its side.

Four tilt switches placed in a cross-shaped pattern on a flexible mount can be used as the basis of a very simple game controller, with a joystick mounted in the center.

Variants

The three configurations of ball-type tilt switches shown in [Figure 8-3](#) are functionally interchangeable and can be chosen for convenience of their leads in fitting the circuit.

Mercury Switches

Early tilt switches contained a blob of mercury in a glass bulb. When the bulb was tilted, the mercury rolled to the end and made an electric

connection between two metal contacts that penetrated the bulb.

A small mercury switch is shown in [Figure 8-6](#). This type of sensor became less common after many countries classified mercury as an environmental hazard and established regulations restricting its use.

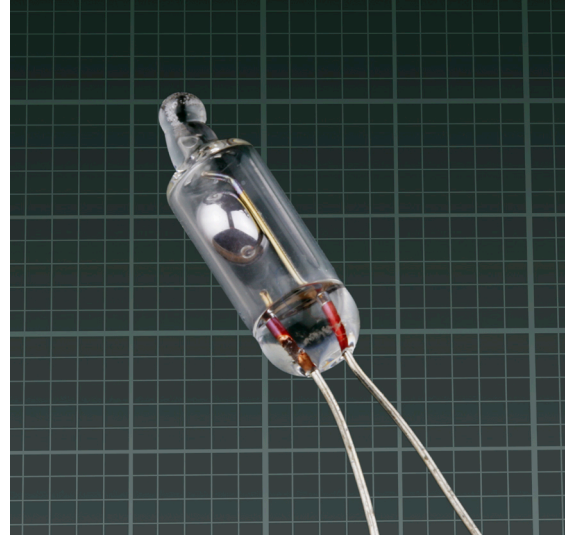


Figure 8-6 Small mercury switch rated for 0.3A at 24VDC or 24VAC. Larger mercury switches can switch more power; 1A at 230V is common. The background grid is in millimeters.

Mercury is an excellent electrical conductor. It remains in a liquid state between about -38 degrees Celsius and +356 degrees Celsius, and has very high surface tension, encouraging it to form a single blob instead of breaking up into small droplets. Because free space in the bulb is filled with an inert gas to prevent oxidation of the electrodes, a mercury switch can have a very long operating life. In the United States in the 1970s, some light switches were sold containing mercury switches with a claimed life expectancy of 100 years.

Pendulum Switch

This type of switch, now relatively rare, was found in vintage pinball machines. It consists of a pendulum about 5cm long, suspended inside

a steel ring about 1cm in internal diameter. If the machine was rocked sufficiently to bring the pendulum in contact with the ring, the game was cancelled and the word “Tilt” appeared on the display. Consequently this was referred to as a tilt switch, although really it was a form of **vibration** sensor with a long period of oscillation.

Magnetization

Some tilt switches use a steel ball that is weakly magnetized, so that it will seat itself more firmly when it rolls into a circular depression or ring. This type of switch must be tilted back through a larger angle to dislodge the ball. Therefore it will exhibit greater *hysteresis*.

Tilt Sensors

Unlike a tilt switch, a tilt sensor is not built around an electromechanical switch.

The principle of a rolling ball has been miniaturized and encapsulated in a small enclosure (10mm square or smaller), in which the ball rolls to interrupt a beam from an internal LED shining on a phototransistor. Examples are found in the Panasonic AHF series. Internal circuitry ensures a clean on-off signal, free from the switch bounce that is a problem in basic ball-type tilt switches. However, the switch requires a power supply, and the open-collector output must be used with a pullup resistor. By comparison, a simple electromechanical tilt switch can be wired directly to the device that it controls.

Diagrams from the Panasonic datasheet show the three types of AHF sensor available for vertical, horizontal, and reverse mounting. In each case, the ball (dotted circle) rests in a shallow cup (dotted curve) where it obstructs the beam from an internal LED (not shown). See [Figure 8-7](#). An exterior view of the AHF22 is shown in [Figure 8-8](#).

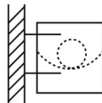
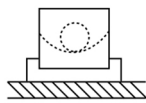
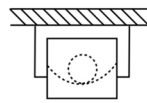
Vertical mounting	Horizontal mounting	Reverse mounting
AHF21	AHF22	AHF23
		

Figure 8-7 Three types of Panasonic tilt sensor, from the manufacturer’s datasheet. See text for details.



Figure 8-8 Exterior view of the Panasonic AHF22 tilt sensor. The background grid is in millimeters.

Two-Axis Tilt Sensors

The Rohm RPI-1035 is a surface-mount tilt sensor about 4mm square, with two phototransistor outputs that indicate which axis the sensor is tilting around. The outputs can be interpreted as a 2-bit binary number, with its four possible states indicating the rotation of the switch around two axes at 90 degrees to each other. Switches of this type were developed to indicate the orientation of consumer-electronic devices such as digital cameras, but more sophisticated sensors containing accelerometers are becoming price-competitive.

Surface-mount 2-axis tilt sensors have been made available on small breakout boards that are easy to use experimentally. An example is the Parallax 28036 shown in [Figure 8-9](#).



Figure 8-9 A 2-axis tilt sensor mounted on a breakout board available from Parallax.

The behavior of the rolling-ball sensor at the heart of this board is shown in [Figure 8-10](#). The sensor contains a square cavity in which the ball is depicted as a blue circle. At one corner of the cavity is a red LED, while two phototransistors, labeled A and B, are at the corners on the left and right. When the sensor is held as in section 1 of the figure, with the LED at the top and the ball resting at the bottom, both phototransistors have a high output as they receive light from the LED.

In section 2 of [Figure 8-10](#), the sensor has been turned through 90 degrees. The ball now prevents light from reaching sensor B, while phototransistor A is still active. In section 3 of the figure, the sensor has been turned through another 90 degrees, so that the ball prevents light from escaping from the LED, and both phototransistors are now dark. In section 4, the ball obstructs phototransistor A but not phototransistor B.

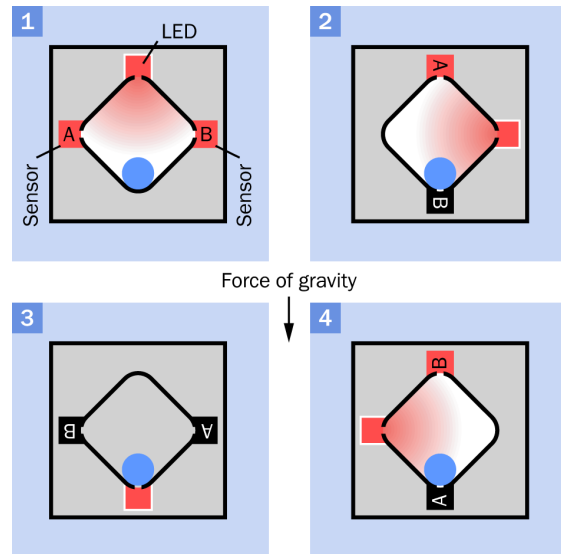


Figure 8-10 The rolling-ball sensor inside the Parallax 28036. See text for details.

Suppose that the sensor is placed flat on a horizontal surface. The sensor will now respond to being tipped either way around two horizontal axes. This justifies its description as a 4-directional tilt sensor, although its design suggests that it may have been intended for use as described above, rotating through four positions around one horizontal axis.

Values

A heavy-duty tilt switch can be rated as highly as 10A at 240VAC. More commonly, a tilt switch about 15mm long can be expected to switch about 0.3A at 24VAC or 24VDC.

The *operating angle* is the angle through which the switch must be turned to activate it, relative to its normal rest position.

The *return angle* is the angle to which the switch must be returned to deactivate it. Hysteresis results from the return angle being smaller than the angle that activates the sensor.

Tilt sensors with an open-collector output will specify maximum forward current for the internally mounted LED (usually no greater than

50mA), the maximum collector-emitter voltage at the output (typically 30V), and the maximum collector current (often 30mA). An explanation of open-collector outputs is given in [Appendix A](#) (see “3. Analog: Open Collector”).

How to Use It

An electromechanical tilt switch can be connected directly between a power supply and a device, so long as the device does not draw more current than the switch is rated to handle. Note that inductive loads such as motors draw an initial surge that can be at least twice the operational rating, while relays may be more likely to create a voltage spike when disconnecting. Switches should be chosen accordingly. For more discussion of this topic, see **switch** in Volume 1.

A small tilt switch can be used in conjunction with a relay or transistor to amplify its signal sufficiently to drive a larger load.

If an electromechanical tilt switch is connected to an electronic device such as a microcontroller or logic chip, output from the switch will have to be *debounced* to prevent a series of brief voltage spikes that can cause false triggering when the switch turns on or off. A debouncing logic circuit or chip can be used, or the program code in a microcontroller can introduce a wait period of up to 50 milliseconds to allow the contacts to settle.

Mercury switches are much less likely to create a noisy output than rolling-ball switches, and may require little or no debouncing.

For an application that must sense rotation around two or three axes, multiple single-axis tilt switches can be combined. A microcontroller or logic gates will be necessary to evaluate signals from the switches, to determine the orientation.

What Can Go Wrong

Contact Erosion

If a ball-type tilt switch is subjected to current that exceeds its specification, arcing may erode its contacts, and they will become less reliable, especially if the contacts are plated with a thin metallic film that is eroded. For additional information on arcing in switches, see the **switch** entry in Volume 1.

Random Signals

During the brief time when a ball-type tilt switch is turning from one position to the other, vibration of the ball(s) inside it is likely to create erratic, random signals. If the output from the switch is being evaluated by a microcontroller, a debouncing routine may be insufficient to prevent the random signals from being sensed, and some programming will be necessary to ignore the signals during this transitional phase. If the switch is connected directly to a relay, the intermittent signals may occur sufficiently rapidly that the relay will ignore them.

Environmental Hazard

A device that incorporates a mercury switch may have to be re-engineered in the future if the availability of mercury switches becomes unreliable as a result of tighter environmental regulations. For the same reason, the end user may have difficulty replacing a mercury switch if it fails. Therefore, a ball-type tilt switch should be used instead of a mercury switch in any newly designed device.

Requirement for Gravity

Because a tilt switch depends on gravity to roll a ball or move a blob of mercury, it will not work in low-gravity, reversed-gravity, or zero-gravity conditions—for example, in a rocket during the unpowered phase of ascent and descent, or in an aircraft that performs aerobatic maneuvers. Performance of a tilt switch in a vehicle that accelerates or decelerates sud-

denly may also be unreliable. Likewise, it cannot be used on a small boat.

Requirement for Stability

A tilt switch will tend to give erroneous results in a location where there is significant vibration

or where the object containing the switch may be turned or repositioned unpredictably by the user.

gyroscope



Historically, a **gyroscope** always contained a spinning disc. Some devices for navigation still depend on rotating elements, but they are outside the scope of this Encyclopedia. This entry deals primarily with *vibrating gyroscopes*, also known as *resonator gyroscopes*, that are MEMS devices contained within silicon chips.

OTHER RELATED COMPONENTS

- **accelerometer** (see [Chapter 10](#))
- **GPS** (see [Chapter 1](#))
- **magnetometer** (see [Chapter 2](#))

What It Does

A **gyroscope** resists rotation around any axis at right angles to its own axis of rotation or vibration. Consequently, if the gyroscope is allowed to move freely on gimbals in a sealed enclosure, the gyroscope will tend to maintain its orientation while the enclosure can rotate freely around it.

Taking this concept a step further, if the enclosure is mounted in an aircraft, the aircraft's rotation around two axes can be determined by referring to the gyroscope. If additional gyroscopes are added orthogonally to the first, the aircraft's rotation around all three axes can be determined.

A gyroscope does not measure linear motion in any direction, or any static angle of orientation.

Schematic Symbol

A chip-based gyroscope, magnetometer, or accelerometer may be represented in a schematic as a rectangular box containing abbreviations to

identify pin functions (as in any integrated circuit chip).

IMU

An **accelerometer** measures variations in linear motion and will also measure its own static orientation relative to the force of gravity. If an accelerometer rotates around its own axis, it will not measure angular velocity.

A **magnetometer** measures the magnetic field surrounding it, and may be sufficiently sensitive to determine its orientation relative to the Earth's magnetic field.

When an accelerometer and a gyroscope are contained in the same package, optionally with a magnetometer, they may be described as an **IMU** (inertial measurement unit), which can provide necessary data to maneuver aircraft, spacecraft, and watercraft, especially when **GPS** signals are unavailable.

Applications

The first chip-based gyroscope was used in automobiles in 1998 as a yaw sensor in a skid-

control system. Subsequent automotive applications include active suspension control, air bag sensors, rollover detection and prevention, and navigation systems.

Gyroscopes may be installed in military ordinance to provide backup in case an onboard **GPS** system fails, possibly as a result of radio jamming.

Handheld 3D game controllers and headsets may use gyroscopes to control images displayed to the viewer. A digital camera may employ a gyroscope to provide image stabilization. Gyroscopes are usually found in quadcopters or drones, are used to stabilize two-wheeled vehicles such as the Segway, and are used in robotics.

How It Works

The traditional form of gyroscope is a rotating wheel, which will resist turning forces perpendicular to its own axis of rotation. In [Figure 9-1](#), three directions at right angles to each other are defined in the bottom-right corner of the diagram as X, Y, and Z. The wheel is rotating around the X axis, as shown by the green arrow. It will resist any turning force around the Y axis (red arrows) or the Z axis (yellow arrows).

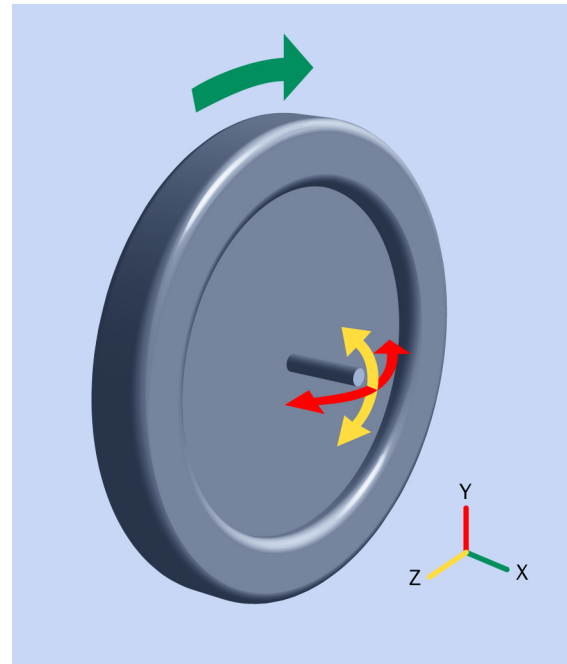


Figure 9-1 In a traditional gyroscope, a wheel that is rotating (shown by the green arrow) will resist a turning force on either of the axes (shown by red and yellow arrows) perpendicular to its axis of rotation.

Vibrating Gyroscope

A vibrating fork can be substituted for a wheel. In [Figure 9-2](#) a fork is secured at its base while its tines are induced to vibrate toward each other and away from each other, as suggested by the double-ended arrow. In a chip-based gyroscope, this vibration is induced piezoelectrically or by static electricity.

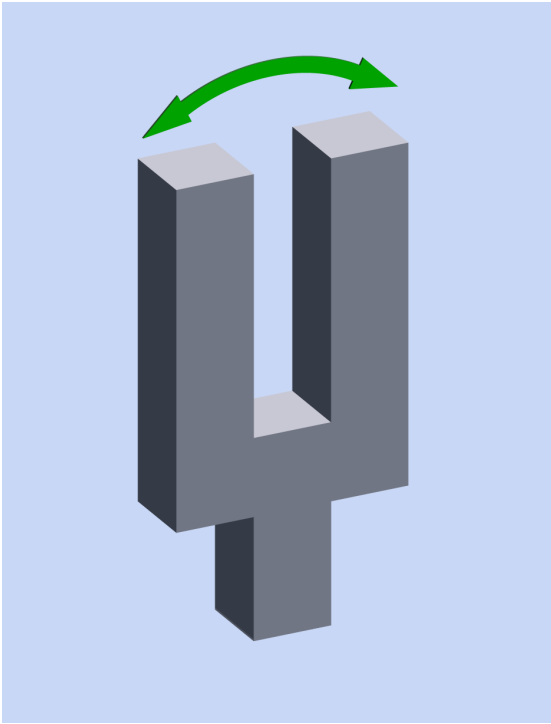


Figure 9-2 A vibrating fork (green arrows) can be substituted for a rotating wheel in a gyroscope.

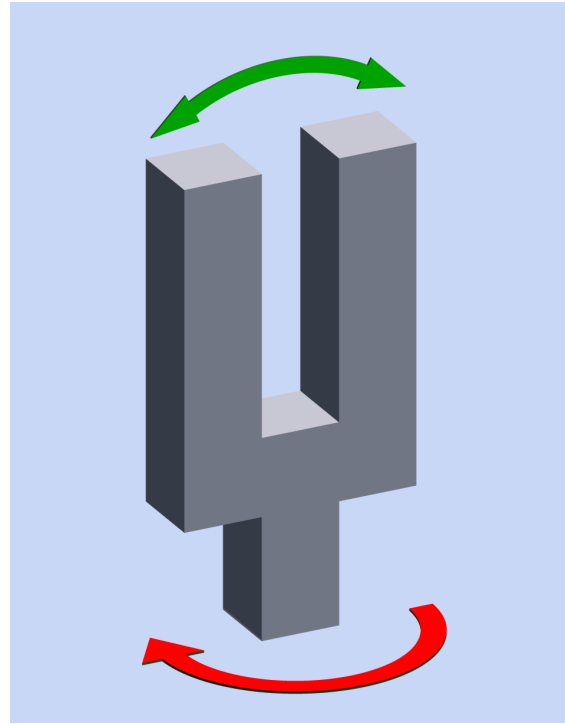


Figure 9-3 A turning force is applied to the base of the fork around a vertical axis, as shown by the lower arrow.

Now suppose a turning force is applied around the vertical axis at the base of the fork, as suggested by the lower arrow in [Figure 9-3](#).

The angular momentum of the vibrating tines causes them to resist this turning force, and consequently they will tend to bend, as shown by the yellow arrows in [Figure 9-4](#). The amount of their deflection can be measured capacitively. This system is used in many chip-based gyroscope systems, and may be referred to as a [vibrating gyroscope](#) or a [resonator gyroscope](#).

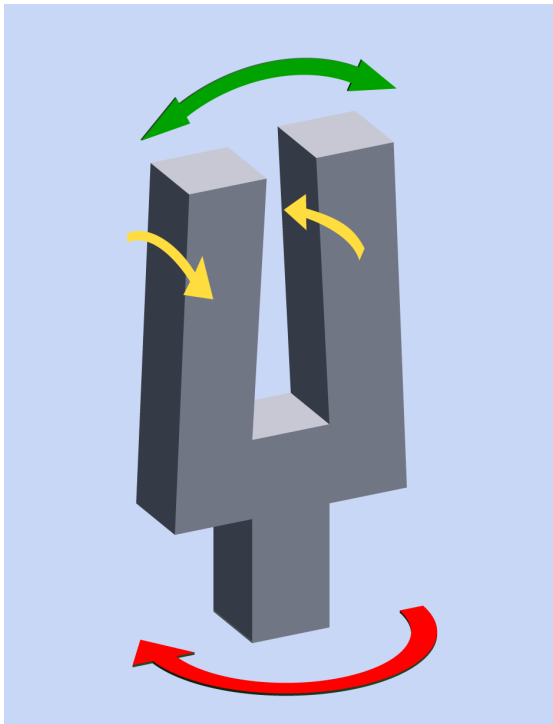


Figure 9-4 The angular velocity of the rotating fork causes deflection of its vibrating tines, shown by the yellow arrows.

An assembly of microscopic forks can be etched into a silicon chip. In [Figure 9-5](#), an electron micrograph shows the interior of this type of chip. It contains three rotational sensors, responding to the X, Y, and Z axes of motion. These sensors respond to *pitch* (rotation around the X axis), *roll* (rotation around the Y axis), and *yaw* (rotation around the Z axis).

Fork-based sensors are analog devices whose values are converted to digital values by an onboard *analog-to-digital converter* (ADC). The values are stored in registers that are available to other devices, often via the I2C protocol, which is widely used by microcontrollers.

For additional details about I2C, see [Appendix A](#).

Typically there will be two 8-bit registers for each axis. Each register stores the binary equivalent of a signed integer, where the positive or

negative value represents the direction and magnitude of deflection, usually in degrees per second (dps).

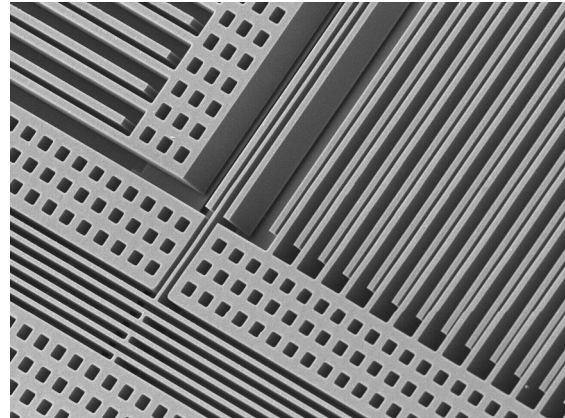


Figure 9-5 Electron micrograph of the STMicroelectronics LIS331DLH vibrating gyroscope installed in the Apple iPhone 4. The parallel plates at bottom-left function as a spring, while the elements at top-left and at right measure capacitance as their orientation varies according to rotational velocity. Photo courtesy of MEMS Journal published by *Chipworks*

Variants

The L3G420D by STMicroelectronics is a 3-axis gyroscope-only chip. It communicates via the SPI or I2C protocol, is approximately 4mm square, and can measure rotational rates up to plus-or-minus 2,000 degrees per second.

The Freescale FXAS21002C has a similar specification. Gyroscope-only chips of this type have fallen in price to the point where they are comparable to the retail cost of everyday components such as a small-signal relay or an audio amplifier on a chip.

IMUs

Chips that only contain gyroscopes are becoming less common as the cost of adding accelerometers decreases.

Gyroscopes and accelerometers are complementary, as gyroscopes are insensitive to linear motion or the Earth's gravity, but accelerome-

ters can measure the rate of change of linear motion and orientation of the chip relative to the Earth. Software can combine this data to calculate the shape of the path being described by a device containing the chip, in addition to the changing velocity of the chip along that path.

The InvenSense MPU-6050 is a common 3-gyroscope, 3-accelerometer chip. It also includes an interface for connecting an external 3-axis magnetometer. The SPI and I2C communications protocols are supported. The MPU-6050 has been a popular choice in the hobby-electronics community, so that Arduino-compatible code to interpret its data is available from many sources. Breakout boards are available with the MPU-6050 installed. An example is the Sparkfun SEN-11028 shown in Figure 9-6.

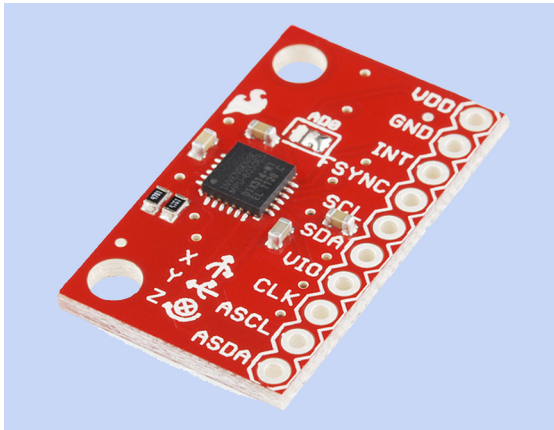


Figure 9-6 A breakout board from Sparkfun, providing easy access to the InvenSense MPU-6050 chip combining three gyroscopes and three accelerometers.

Values

The *rotational velocity* of a gyroscope element is usually expressed in degrees of rotation per second (dps), and sometimes in rotations per minute (RPM).

A datasheet will specify the number of sensor axes (usually 3), supply voltage (3.3VDC is com-

mon), maximum digital-low and minimum digital-high output voltages, and power consumption in normal mode and sleep mode. Power consumption is usually less than 10mA.

The *dynamic range* is the maximum forward and reverse rotational velocity, which usually will not exceed plus-or-minus 2,000 degrees per second. Lower ranges may be user-selectable. The advantage of selecting a lower maximum rate of change is that it can be converted to a digital value with higher precision.

The *sensor resonant frequency* will be several kilohertz, and must be higher than the frequency of any vibration that is applied to the sensor during use.

The *communications protocol* is usually I2C and SPI may be offered as an option, with a selectable digital output data rate.

Bias temperature coefficient describes the effect of temperature on the gyroscope.

The *resolution* of the gyroscope relates to the number of bits used in the digital output from the onboard ADC. A 16-bit resolution is common.

How to Use It

Using a smart chip such as the MPU-6050, the circuit designer can take advantage of its onboard digital motion processor (DMP). Still, obtaining orientation information from the contents of registers on the MPU-6050 is non-trivial. Online sources and code libraries are necessary. The book *Make: Sensors* contains code listings for the Raspberry Pi as well as the Arduino.

What Can Go Wrong

Temperature Drift

Vibrating materials at the heart of a chip-based gyroscope are likely to change their behavior with temperature. Typically the chip will con-

tain a temperature sensor, the value of which can be used to adjust the output value of the gyroscope.

Mechanical Stress

Stress can be induced when a surface-mount chip is soldered to a board. The vibrating parts of a chip-based gyroscope may be adversely affected. Datasheets will supply information regarding maximum acceptable temperature during the soldering process.

Vibration

Because a chip-based gyroscope depends on the consistent behavior of internal vibrating

parts, external vibration can degrade its accuracy. Sensor design can minimize the effects of vibration, but the datasheet should be consulted for details.

Placement

A gyroscope should be placed on a circuit board near a hard mounting point where deflection of the board will be minimized.

accelerometer

10

OTHER RELATED COMPONENTS

- **GPS** (see [Chapter 1](#))
- **gyroscope** (see [Chapter 9](#))
- **tilt** sensor (see [Chapter 8](#))
- **vibration** sensor (see [Chapter 11](#))

What It Does

Acceleration is the rate at which velocity changes over time. If a car takes 10 seconds to increase its speed from 30kph to 40kph relative to the road on which it is traveling, it is accelerating at an average of 1kph each second. If it then reduces its speed back to 30kph during another 10-second interval, it is decelerating at the same rate—although deceleration is really just negative acceleration.

While a car is accelerating, passengers will feel a lateral force exerted on them. Similarly, astronauts in a rocket that blasts off will feel a downward force. According to Einstein's theory of equivalence, forces resulting from acceleration are indistinguishable from the force of gravity.

Consequently, a sensor that measures acceleration can also measure gravity. This sensor is an **accelerometer**. Its output may be measured in *gravities*, abbreviated with the letter g (not to be confused with the usage of G to measure the strength of a magnetic field in gauss).

If three accelerometers are mounted orthogonally (at 90 degrees to each other), their readings can reveal:

- The direction of acceleration of a moving object.
- If an object has been dropped and is falling freely.
- Which way up it is being held in a stationary position.
- The severity of an impact when a moving object collides with some other object.

IMU

A **gyroscope** measures the rate of rotation of the enclosure in which it is mounted. This is properly known as the *angular velocity*. A gyroscope will also respond to changes in the rate of rotation. It does not measure linear motion or a static angle of orientation.

A **magnetometer** measures the magnetic field surrounding it, and may be sufficiently sensitive to determine its orientation relative to the Earth's magnetic field.

When an accelerometer and a gyroscope are contained in the same package, optionally with a magnetometer, they may be described as an *IMU* (inertial measurement unit), which can pro-

vide necessary data to maneuver aircraft, spacecraft, and watercraft, especially when **GPS** signals are unavailable.

Schematic Symbols

A chip-based accelerometer may be represented in a schematic as a rectangular box containing abbreviations to identify pin functions, as in any integrated circuit chip. No specific symbol is used for any of these components.

Applications

In the past, accelerometers were laboratory devices that calibrated the performance of cars, airplanes, and other types of vehicles. Measuring the ability of car tires to withstand cornering forces was an application where an accelerometer was used.

Miniaturization of accelerometer elements, coupled with a radical reduction in their cost, has enabled them to be installed in small electronic devices ranging from smartphones to hard drives.

In a phone or a camera, accelerometers can determine which way up the user is holding the device. The camera can rotate the picture appropriately, and the orientation of the picture can be saved with its image data.

In an external hard drive containing rotating platters, accelerometers can protect the read-write heads by rapidly parking them during the fraction of the second that elapses between someone dropping the hard drive and its impact with the floor.

Accelerometers can be installed in a 3D mouse or virtual-reality headset to determine its orientation and motion. This enables a video image to be updated appropriately. For example, the Nintendo Wii Remote has been marketed with an ADXL330 accelerometer.

In an automobile, an accelerometer can trigger the deployment of an air bag when the deceleration caused by an impact exceeds a threshold level.

How It Works

The simplest conceptual model of an accelerometer consists of a mass attached to one end of a coiled compression spring. The other end of the spring is anchored in an object whose acceleration is being measured. The mass can only move along the same axis as the spring.

Figure 10-1 shows three views of a simplified accelerometer, which is a sealed tube shown in dark red. In the center image, the accelerometer is in its rest state. The top image shows the mass (a dark blue square) responding when the tube accelerates from left to right. The third image shows it decelerating (that is, undergoing negative acceleration, or acceleration from right to left). Using an ideal spring, the displacement of the mass will be proportional to the rate of acceleration, within reasonable limits. The displacement can be measured optically or capacitively.

Note that the rest state will resume when the acceleration stops, regardless of constant motion in any direction. An accelerometer only measures a *change* in velocity. It does not measure a constant velocity.

An accelerometer cannot measure rotation around its own axis of movement. Therefore, it may be used in conjunction with a **gyroscope**, which measures angular velocity.

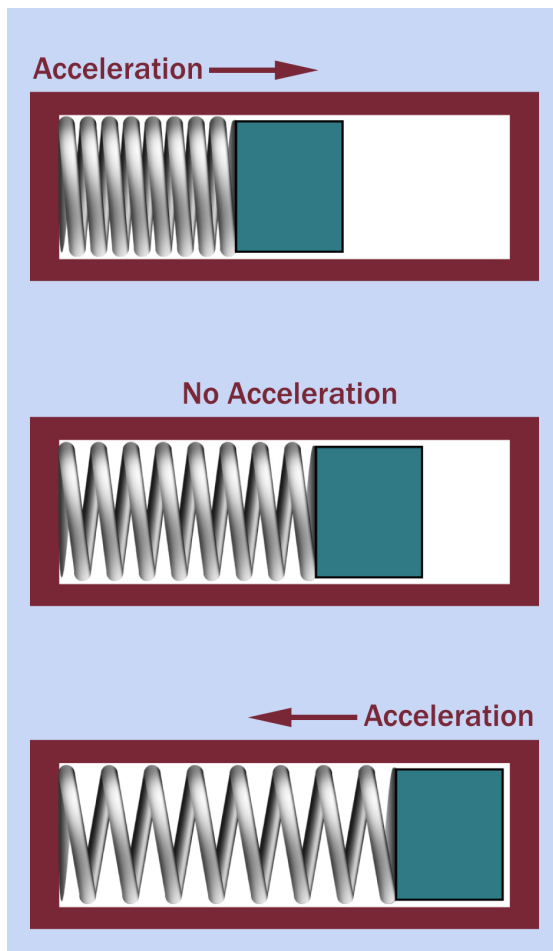


Figure 10-1 Simplified view of an accelerometer consisting of a sealed tube (dark red outline) with a spring anchored to it at left. A small mass (dark blue) is attached to the right end of the spring. The mass responds to acceleration of the tube.

Gravity and Free Fall

If a device containing the simplified accelerometer described previously rests on the ground, and is mounted vertically as in the left section of [Figure 10-2](#), gravity acting on the mass will apply a force to one end of the spring, while the other end is restrained. The accelerometer will now measure 1g as a downward force.

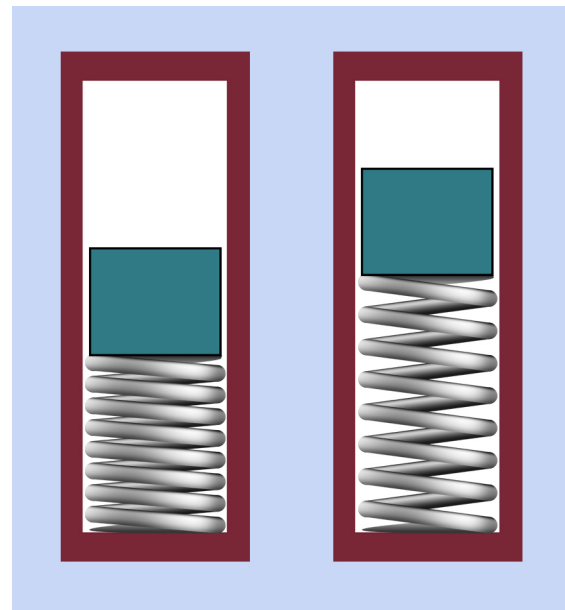


Figure 10-2 Left: an accelerometer resting on the ground measures 1g, as the force of gravity pulls the mass downward. Right: in free fall, the accelerometer measures 0g.

If the device is dropped so that it falls freely under the influence of gravity, it is in *free fall* and will accelerate at approximately 9.8 meters per second each second. This is usually written as 9.8m/sec^2 , or can be described as 1 gravity, often expressed as 1g.

An accelerometer in free fall will measure 0g, as shown on the right in [Figure 10-2](#), because all parts of the accelerometer are now accelerating equally under the force of gravity.

If three accelerometers are assembled orthogonally, and if one accelerometer is vertical, and the device is held motionless relative to the Earth, the vertical sensor will show 1g while the other two accelerometers will show 0g. If the device is dropped, all sensors will show 0g.

Rotation

If a device containing three accelerometers is turned over, and is not in free fall, accelerometers mounted orthogonally will show values that vary depending on their alignment with the force of gravity.

Calculation

The force acting upon an object will cause a rate of acceleration that can be calculated by Newton's Second Law of Motion, provided that the object can move freely and is not being subjected to any additional forces, such as the force of gravity. If F is the force, m is the mass of the object, and a is the acceleration:

$$F = m * a$$

And therefore:

$$a = F / m$$

If the mass is restrained by a spring whose compression or extension has an approximately linear relationship to the force applied to it, acceleration can be calculated as a function of the linear displacement of the mass.

These statements ignore relativistic effects that are insignificant unless ultra precise time and motion measurement may be involved.

In a real-world accelerometer, movement of the mass will require some form of damping to prevent it from oscillating.

Variants

Accelerometer prices dropped radically after 2010. In an effort to maintain profitability, manufacturers have loaded more features onto chips. While a 2-axis accelerometer such as the Memsic 2125 seemed a good option when first introduced, it is now facing obsolescence as 3-axis accelerometer chips that also contain 3-axis gyroscopes have become ubiquitous—and no more expensive.

Early chip-based accelerometers provided analog outputs where voltage was proportional with acceleration and could be processed by a comparator. On some breakout boards, such as the Dimension Engineering DE-ACCM6G, which used the STMicroelectronics LIS244ALH chip, a comparator was included (see [Figure 10-3](#)).

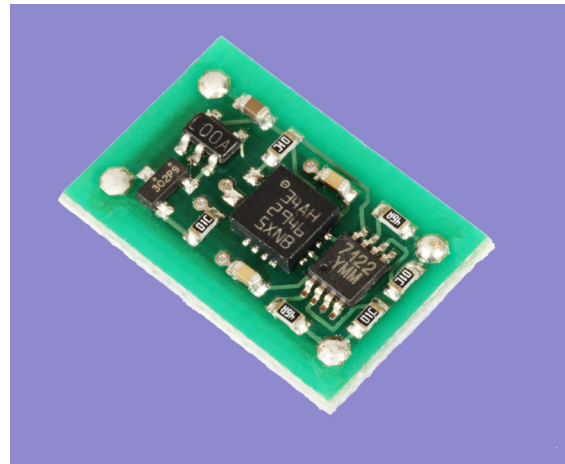


Figure 10-3 A relatively early, relatively expensive breakout board using the 2-axis accelerometer LIS244ALH chip by STMicroelectronics. It has been superseded by chips that combine accelerometers, gyroscopes, and processors to provide a digital output.

Because this board allowed a maximum output of only 0.83mA, it was only suitable for high-impedance logic chips or a microcontroller. However, because the output was analog, it could be passed through another comparator for direct connection to a piezo beeper to create a device that would sound an alarm when tilted. This is shown in [Figure 10-4](#).

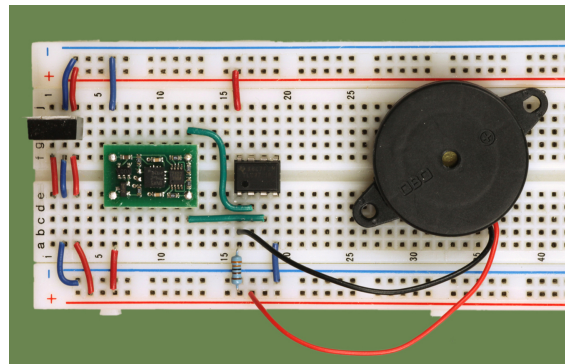


Figure 10-4 Connecting the analog output from the DE-ACCM6G 2-axis comparator breakout board through a comparator to a piezo beeper.

Many chips now contain gyroscopes as well as accelerometers. The electron micrograph in [Figure 10-5](#) shows the interior of a chip of this

kind, where the zig-zag shapes are “springs” etched into the die, the large areas patterned with dots are masses that can respond to various forms of motion, and the parallel plates are capacitive sensors.

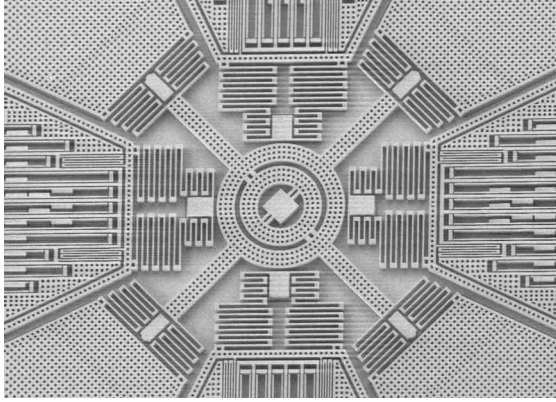


Figure 10-5 Interior of a chip combining gyroscopes with accelerometers.

The complexity of a chip combining two types of sensors creates a need for more complicated code to process the six outputs. Consequently most accelerometer chips now have their own onboard ADCs, and digital registers for communication with microcontrollers via the I2C communication protocol. Some chips also have onboard processing to interpret the mix of data. However, microcontrollers still need code to make sense of the data.

The hobby-electronics community has responded. A breakout board such as the LSM9DS0 from Adafruit, using a chip from STMicroelectronics that shares that same part number, tries to make this extremely complex chip accessible to experimenters. Shown in [Figure 10-6](#), the LSM9DS0 contains a 3-axis magnetometer in addition to a 3-axis gyroscope and a 3-axis accelerometer.

The datasheet for the LSM9DS0 runs to more than 70 pages. At the time of writing, Adafruit is still refining code for the Arduino to make the features of this chip accessible.

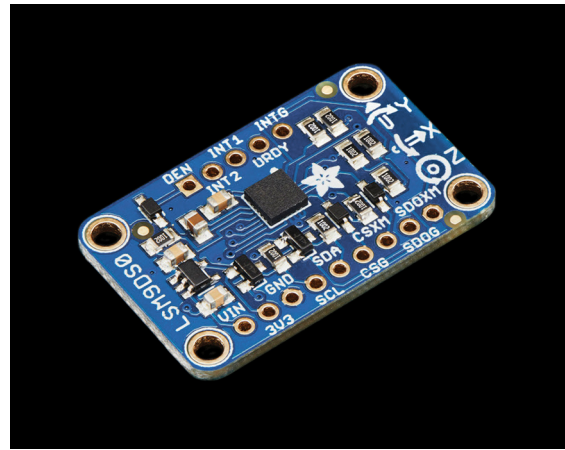


Figure 10-6 This breakout board from Adafruit is built around an LSM9DS0 chip that combines accelerometers with gyroscopes and magnetometers.

Despite the complexity of the LSM9DS0 chip, the breakout board sells for approximately the same price, at the time of writing, as the DE-ACCM6G 2-axis analog-output magnetometer four years ago. In the future we may expect IMU prices to fall still further, so that this type of multifunction chip becomes the default, and users simply ignore the functions that they don’t need.

Values

Current consumption will vary depending on the activity of the chip, and may be broken down for different types of sensors on multi-sensor chips. A modern accelerometer may typically draw less than 1mA. Gyroscope power consumption will be greater, as a segment of the chip will be maintained in a state of vibration.

Linear acceleration measurable by an accelerometer is customarily expressed in gravities, abbreviated g. Current generations of chips may be able to measure as much as plus-or-minus 16g, but because the value is converted internally to a digital quantity, smaller accelerations will not be expressed so accurately. Therefore the measurement range for acceleration is

usually user-selectable by sending an appropriate code to the chip. Ranges may include plus-or-minus 2g, 4g, 6g, 8g, and 16g.

Sensitivity defines the smallest increment of acceleration measurable by the least significant bit (LSB) in an output register, for each of the acceleration ranges. In a range of plus-or-minus 2g, the internal 16-bit ADC may be capable of measuring 0.06 milligravities. When the range is reset to plus-or-minus 16g, the smallest increment may be around 0.7 milligravities.

The measurable range of gravities should not be confused with the maximum acceleration that the chip can tolerate without suffering internal mechanical damage, either while it is powered or unpowered. This will be more than 1,000g provided the duration is brief, and would only be experienced during an impact.

Linear acceleration sensitivity change versus temperature is usually expressed as a percentage, such as plus-or-minus 1.5%.

Output type will be analog or digital. If digital, the data protocol will be I2C or SPI. If I2C, the

address of the device should be configurable. The data transfer rate is typically at least 100kHz, and this also may be configurable to different rates.

For additional details about protocols such as I2C, see [Appendix A](#).

What Can Go Wrong

Mechanical Stress

Stress can be induced when a surface-mount chip is soldered to a board. The moving parts in a chip-based accelerometer may be adversely affected. Datasheets will supply information regarding maximum acceptable temperature during the soldering process.

Other Problems

If the accelerometer function is combined with other sensing functions such as magnetometer or gyroscope, see [Chapter 2](#) or [Chapter 9](#) for additional cautions about potential problems affecting these kinds of sensors.

vibration sensor

11

An **accelerometer** can measure some aspects of vibration. However, this entry deals primarily with mechanical devices (often described as *vibration switches*) and piezoelectric devices (often described as *vibration sensors*) that are solely intended to measure vibration.

A *vibrometer* measures vibration using a laser beam aimed at a reflective spot applied to a surface. It is usually a laboratory instrument, beyond the scope of this Encyclopedia.

OTHER RELATED COMPONENTS

- **accelerometer** (see [Chapter 10](#))
- **tilt** sensor (see [Chapter 8](#))
- **force** sensor (see [Chapter 12](#))

What It Does

A vibration sensor responds to repetitive mechanical motion. Most versions contain two switch contacts that are normally open and will close if the sensor vibrates in its designed frequency range. In some sensors, the frequency range and sensitivity are manually adjustable.

Large sensors are used as automatic shutdown switches responding to excessive vibration in machinery, and may be capable of switching substantial currents (10A or higher). Smaller versions can shut down domestic appliances such as a washing machine that is seriously out of balance during a spin cycle.

A vibration sensor can be used as a simple user-input device in toys and games.

A shock sensor can detect abuse of a sensitive device, for example, by including the sensor and a data logger when the device is transported.

Schematic Symbols

Either of the symbols in [Figure 11-1](#) may represent a piezoelectric or piezoresistive vibration sensor, but they also represent other piezo-based devices.

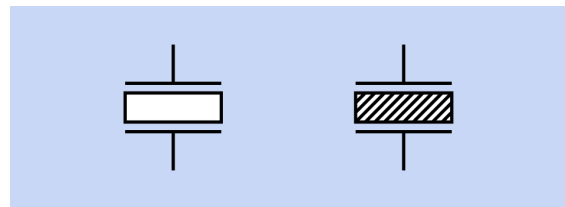


Figure 11-1 Either of these symbols may represent a piezoelectric or piezoresistive device, including (but not limited to) vibration sensors that operate on this principle. The symbol on the left is more common.

Variants

Vibration sensors use a wide variety of detection methods.

Pin-and-Spring

Probably the simplest type of sensor consists of a small, thin pin in the center of a miniature coil spring. The spring is anchored at its base, but its other end is free to vibrate. If the vibration reaches a sufficient amplitude, the spring touches the pin, completing a circuit between the two leads of the device.

An example is shown in [Figure 11-2](#), where two identical sensors are shown, one of them cut open to reveal the gold-plated rod and spring inside. This sensor is rated for 10mA at up to 12VDC.

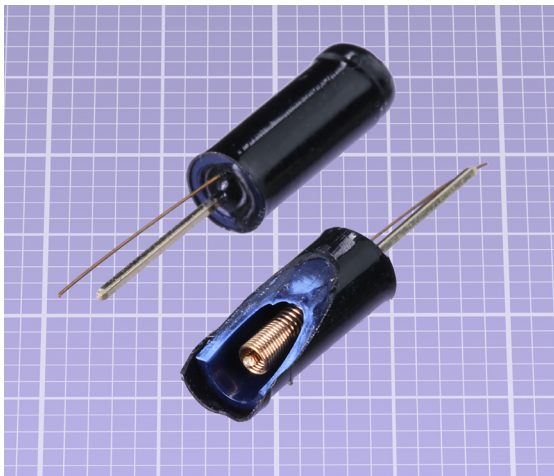


Figure 11-2 When the spring in this sensor vibrates, it touches the pin that is centered in it. The background grid is in millimeters.

Advantages of this system are low cost, ability to respond along two out of three axes, no power supply requirement, and ability to switch AC or DC. However, because the internal contact is extremely brief, it must be connected with a latching component of some type. A flip-flop could be used, or a 555 timer. The switch may also be connected with an input pin on a microcontroller, provided a pullup or pulldown resistor is used to prevent the input from floating when the switch is open.

Externally, the packaging of a pin-and-spring vibration sensor is almost indistinguishable

from a small **tilt** sensor that contains one or two spherical metal balls. The tilt sensor may respond to vibration, but only of a large amplitude and low frequency.

A miniature board containing a pin-and-spring vibration sensor, comparator, and trimmer potentiometer for sensitivity control is sold cheaply by the Chinese supplier Elecrow as their product SW-18015P, shown in [Figure 11-3](#). Elecrow also offers a wide range of other low-cost sensors.

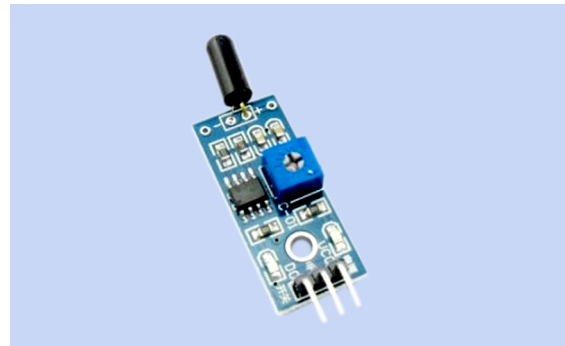


Figure 11-3 A pin-and-spring vibration sensor mounted on a board with sensitivity control.

Piezoelectric Strip

The LDT0-028K by Measurement Specialties is a section of piezoelectric polymer film laminated to a polyester substrate. The film is designed to be anchored at one end, allowing the other end to vibrate. An unweighted version and a weighted version are shown in [Figure 11-4](#), each measuring about 13mm x 25mm. Addition of the weight alters the resonant frequency of the sensor.

Deflection of about 2mm is sufficient to generate a surprising 7VDC between the two leads. Larger deflections will generate higher voltages. The manufacturer suggests that direct connection to a CMOS component is possible. An op-amp may be used for signal conditioning.

A piezoelectric device of this type only generates voltage during the process of deflection. If

the strip is held in a curved position, its output diminishes to zero.

This sensor has a resonant frequency around 170Hz when there is no weight attached to its free end.

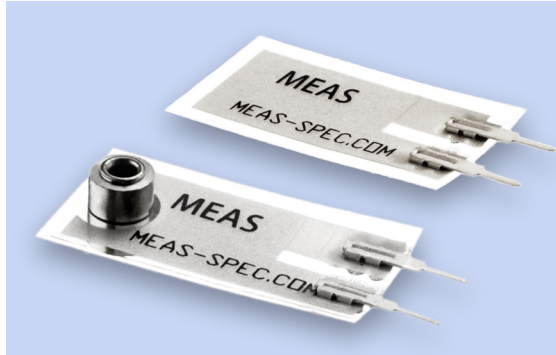


Figure 11-4 Two versions of the LDT0-028K vibration sensor from Measurement Specialties, one with an added weight to lower its resonant frequency.

Chip-Based Piezoelectric

The Murata PKGS series is an example of a surface-mount piezoelectric shock sensor. Measuring only about 1mm x 2mm x 4mm it has an analog output designed for connection through an op-amp. The manufacturer suggests application in a hard-disk drive to block read-write operations when vibration occurs. Similarly, these shock sensors may be used in CD-ROM or DVD drives. They may also be installed in cash dispensing machines to sound an alarm if vandalism occurs.

Toshiba's TB6078FUG is a similar product. Note that because these devices contain electronic components, they require a power supply (usually 3.3VDC to 5VDC) for operation.

“Mousetrap” Type

Some vibration switches rely on a simple system of leverage that is comparable to a mousetrap, in that a relatively small stimulus releases a strong spring. In [Figure 11-5](#) the upper part of the figure shows the switch at rest, held in its position by a powerful spring and by the weight of a mass attached to a pivoted arm. In

the lower part of the figure, severe vertical vibration has caused the assembly to move up and down with sufficient energy to overcome the tension in the spring while the inertia of the mass has resisted the motion. Consequently the arm has moved past the position where the spring is aligned with the pivot, and the spring now acts to hold the arm against a snap-action switch. A system of this type is used in sensors on some power-station cooling towers, where the loss of a large fan blade can result in major vibration.

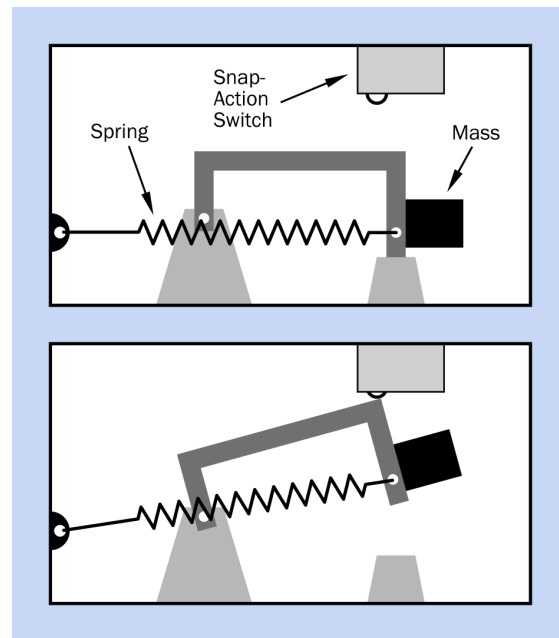


Figure 11-5 A spring-loaded vibration sensor.

Magnetic

Typically this method is used to detect excessive vibration in machines or other devices containing heavy rotating mechanical components. The sensor may be physically large, moving switch contacts that are designed to handle currents of 1A or significantly more.

In one system, a steel ball is retained by a permanent magnet that is barely powerful enough to prevent the ball from falling. Excessive low-frequency vibration will dislodge the ball, which falls and completes a circuit between

two contacts. This activates a relay that powers down the piece of machinery that is vibrating.

The ball may be spaced a small distance from the magnet by a beveled nonmagnetic seat, and the seat may be movable with an external screw. This will adjust the sensitivity of the switch. After the switch has been triggered, it must be reset, which may entail using an external lever to raise the ball back to its location near the magnet.

Another magnetic system is shown in [Figure 11-6](#), where a magnet on a vertical arm can be displaced by horizontal vibration, and an inertial mass on a spring-loaded horizontal rod can also dislodge the magnet in response to vibration along the other two axes.



Figure 11-6 A magnetic vibration switch. See text for details.

Mercury

A small [mercury switch](#) may be used as a vibration sensor, although this application is uncommon. See [Chapter 8](#) for more information about mercury switches.

Values

Measurement of vibration is a complex science of special interest in mechanical design, especially in areas such as the powertrain and suspension geometry in an automobile. Only a few fundamentals will be summarized here.

Primary Variables

The four primary variables in vibration are frequency, displacement, velocity, and acceleration. Frequency describes how rapidly the vibration occurs; displacement describes how far the vibrating object moves in each direction; velocity describes how fast it moves during each cycle; and acceleration describes how rapidly the velocity changes during each cycle. Different types of sensors can be chosen for their responsiveness to each attribute.

[Figure 11-7](#) shows the theoretical relationships between displacement, velocity, and acceleration plotted against the frequency of vibration. The y axis of this graph (the vertical axis) is often labeled “amplitude,” but in reality it is being used to measure three different units, as shown. The curves indicate that if the velocity of vibration remains constant while frequency increases, acceleration must increase as a function of the frequency while displacement decreases. The acronym “rms” denotes that the values are measured as the “root mean square” of their fluctuations.

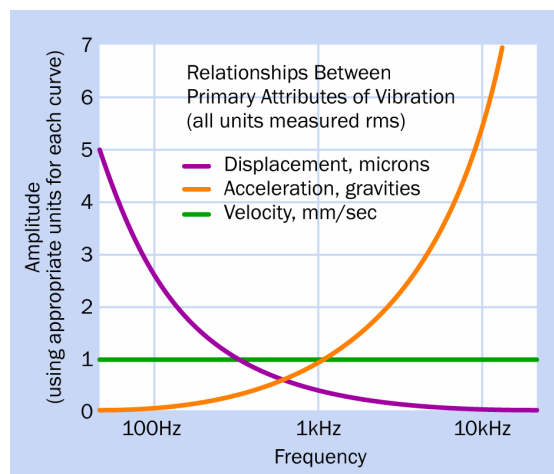


Figure 11-7 Theoretical relationships among the primary attributes of vibration.

Mechanical sensors or switches that respond to displacement are best suited to low frequencies, while piezoelectric sensors that are sensitive to acceleration are best suited to higher

frequencies. Serious mechanical problems tend to result in low-frequency vibrations, while wear on bearings and in gear trains will be more likely to create high-frequency vibration.

Dynamic Attributes

Datasheets for small piezoelectric sensors may only show basic values such as the voltage range that is likely to be created.

Small mechanical sensors of the pin-and-spring type will be rated for maximum voltage and switching current. Typical values may be around 12V (AC or DC) and 10mA, indicating that the output from this type of sensor should be used with an op-amp, microcontroller, logic chip, solid-state relay, or other semiconductor with a high-impedance input.

Sensors that are manufactured for industrial applications will be rated for attributes such as measurable acceleration range (in gravities), temperature sensitivity, frequency response, resonant frequency, capacitance, and power requirement.

The sensitivity of piezoelectric sensors is usually expressed in mV/g. This type of sensor will require a comparator to process the tiny amount of current that it creates.

How to Use It

If a vibration sensor has an analog output requiring a comparator, the output from the comparator is likely to be of the open-collector type. This will require a pullup resistor of a value that provides an appropriate voltage for the next stage in the circuit. For more information about comparators, see Volume 2. For more information about the use of an open-collector output, see “3. Analog: Open Collector”.

A *coupling capacitor* can remove the DC component from the comparator output, allowing only the frequency of vibration to pass through.

The choice of capacitor value will depend on the frequency.

When using a piezoelectric sensor with analog output, a 10M resistor may be installed across its two terminals to reduce voltage drift.

The primary challenge in getting a vibration sensor to work successfully will be matching it to the source of vibration. Manufacturers' datasheets for chip-sized sensors often provide very little information about the optimal values for their products. Peak performance will occur when the natural resonant frequency of a sensor is close to the frequency of vibration that it must detect. Trial and error may be necessary.

A sensor must be mounted appropriately. Most sensors are directional, at least to some extent, and many will not respond significantly to vibration at 90 degrees to their primary axis of sensitivity. Their performance will also diminish if they are placed too far from the vibration source, or if they are mounted on a flexible or yielding surface that will tend to absorb vibration.

While industrial vibration switches may be adjusted manually, the response of small devices designed for circuit-board mounting can only be tweaked using external components to filter out unwanted signals from the sensor.

What Can Go Wrong

Long Cable Runs

The output from a piezoelectric vibration sensor is primarily an AC signal, fluctuating at the frequency of the vibration. Long cabling, or inadequately shielded cabling, can potentially introduce capacitive effects that can degrade the sensor signal. This issue will only affect higher frequencies.

Interference

Sensor signals can also be affected by electromagnetic interference from power lines, transformers, and large motors. This is a significant

issue, as a primary application for industrial vibration sensors is to measure vibration created by motors.

Correct Grounding

For large sensing equipment, grounding may be important to shield cables that transmit data. In an industrial environment, however, grounding is primarily motivated by safety considerations, and an electrical ground can carry

unwanted interference. Ground loops may be created if there are multiple ground points. Ideally, a “ground tree” should be used, where there is only one primary grounding point, and grounds to equipment branch out from it.

Fatigue Failure

In installations where some vibration normally exists, cables should be anchored properly to minimize the risk of fatigue failures.

force sensor

12

A *load cell* or *load sensor* is generally intended to measure a static load, while a *force sensor* can respond dynamically. However, this semantic distinction is not always observed. This entry differentiates between load cells and force sensors but includes them both.

Traditional types of *hydraulic* and *pneumatic* load sensors are not electronic devices. They are outside the scope of this Encyclopedia.

Capacitive, *ultrasonic*, *magnetic*, *optical*, and *electrochemical* force sensors are relatively unusual, and are not included in this Encyclopedia.

A force sensor is occasionally described as a **pressure** sensor, but that term is ambiguous, as it is more often used in conjunction with fluids. This Encyclopedia assumes that a pressure sensor only measures gas or liquid pressure. See Chapter 17.

A **vibration** sensor reacts to rapidly changing forces, but usually cannot measure them accurately, and is simply triggered when vibration exceeds a threshold. See Chapter 11.

Impact sensors that measure the force of a collision are outside the scope of this Encyclopedia.

A sensor designed to respond to a single touch from a fingertip is considered a human-input device and is discussed in the **single touch** sensor entry. See Chapter 13.

OTHER RELATED COMPONENTS

- **vibration** sensor (see [Chapter 11](#))
- **single touch** sensor (see [Chapter 13](#))

What It Does

A **force** sensor measures physical force that is applied to it, either by a person or by an object. Many force sensors respond rapidly and can measure fluctuating forces.

A *load cell* or *load sensor* is usually intended to measure the static weight of an object.

Applications

In robots, force sensors can provide feedback to limit the grip of a mechanical hand. They can also provide haptic feedback for a surgeon performing robotic surgery. In the future, force sensors may find increasing application in agriculture, as mechanized handling of fruit and other foods requires a carefully controlled gripping force.

In medicine, the use of force sensors to evaluate muscle strength in hands or limbs can be important as an indicator of neurological problems or to monitor progress in occupational therapy. Thin force sensors can be installed in shoes to check the weight distribution of each foot. They can also be used for entertainment purposes, to light LEDs in sneakers.

A force sensor may respond to a single-touch user input. See [Chapter 13](#) for more information about **single touch** sensors. Some video game controllers use resistive force sensors to measure the amount of pressure applied to a button. The PlayStation is an example. (Old PlayStation DualShock 2 controllers are a salvageable source of pressure-sensitive buttons.)

Load sensors are used to weigh industrial products, and are also used domestically in kitchen and bathroom scales.

A load sensor can also detect human presence—for example, in the passenger seat of an automobile, where an air bag must not deploy if a young child is present, or in a hospital, to monitor the number of times the patient gets out of bed.

Schematic Symbol

No specific schematic symbol is used for either a force sensor or a load sensor. If a force sensor uses a piezoelectric or piezoresistive element, it may be represented by the symbol shown in [Figure 11-1](#), which is used for many piezo devices.

How It Works

Two methods of force measurement are commonly used: resistive and piezoelectric.

A *piezoelectric* force sensor uses a piezoelectric element, often consisting of a quartz crystal, to convert force to a small voltage that can be amplified. However, this type of sensor only responds to changes in force. If a constant load

is applied, the output peaks quickly and then gradually diminishes to zero.

Resistive force sensors change their electrical resistance when force is applied. They include *metallic strain gauges* and *plastic-film* sensors in which two layers of conductive ink are pressed together.

In SI (standard international) units, the force needed to activate a sensor is measured in newtons, abbreviated with a capital letter N. A newton is defined as the force that would accelerate a mass of 1 kilogram at 1 meter per second each second. More practically, in the gravitational field at the surface of the Earth, 1N = about 100 grams of weight. There are about 28 grams in an ounce; thus 1N is slightly less than 4 ounces.

Strain Gauge

A strain gauge is often made from metallic foil applied to an insulating flexible backing. The backing is glued to a shaped piece of metal, usually steel or aluminum, which is designed to flex slightly under pressure and may be referred to as a *spring*, even though it is often one solid object. Its deflection will be related to the force imposed on it.

The maximum deformation of the spring under a strain gauge is usually 500 to 2,000 parts per million (ppm) when subjected to the maximum force that it is designed to measure. A change of 1ppm is referred to as a *microstrain* (abbreviated $\mu\epsilon$).

The strain gauge has no polarity, and functions like a force-controlled **potentiometer** (see Volume 1). The ratio of the change in its resistance to the change in the strain that it experiences is called the *gauge factor*. For metal foil gauges, the gauge factor is usually around 2.0. This is an approximately linear relationship.

The most common type of foil pattern is shown in [Figure 12-1](#). In the figure, if a stretching force is applied horizontally, the multiple thin sections of foil are slightly elongated, and their

resistance increases within the limits of elasticity of the foil. This effect is multiplied by the number of sections. If the stretching force is applied vertically, the sections are merely separated slightly, and the result is negligible.

Wheatstone Bridge Circuits

The very small changes in resistance in a strain gauge must be amplified to be usable, and the first step is to use a Wheatstone bridge circuit. The simplest form of this circuit is shown in Figure 12-2.

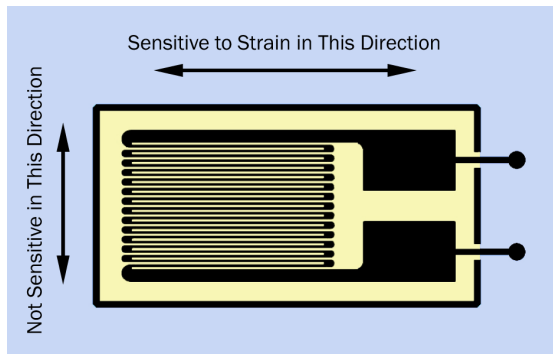


Figure 12-1 The pattern of metallic foil used in a typical strain gauge.

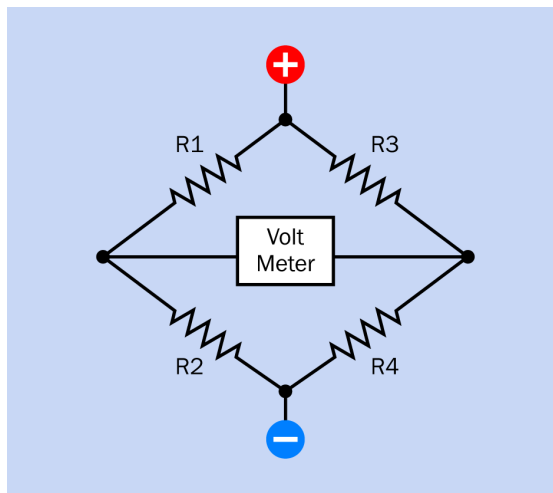


Figure 12-2 A demonstration version of the basic Wheatstone bridge circuit that is often used to detect small changes in a resistance.

Each pair of resistors ($R1 + R2$, and $R3 + R4$) functions as a voltage divider. If all the resistors have an exactly equal value, the voltage at the midpoint of each pair will be identical, and the volt meter at the center will have a zero reading. However, if the value of one resistor changes slightly, the meter will register the imbalance. This circuit is commonly used because of its sensitivity to small variations.

In Figure 12-3 two strain gauges have been substituted for resistors $R3$ and $R4$. The upper strain gauge has been mounted so that it experiences an increase in force at the same time that the lower strain gauge experiences a decrease, usually because one gauge is mounted on the top side of a flexing element while the other gauge is mounted on the underside.

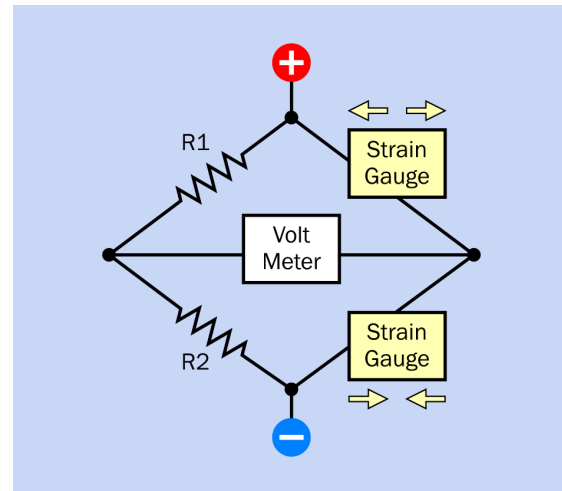


Figure 12-3 Two strain gauges, oppositely oriented, can be used as resistances in the Wheatstone bridge circuit.

Using two strain gauges in this way doubles the sensitivity of the Wheatstone bridge circuit. The configuration is known as a “half Wheatstone” force sensor, and will have three connecting wires. One will be black, one will be red, and the third will be a different color. The red and black wires are for connection to the power supply, as shown in the schematic, while the third wire is common and should be considered as an output.

If an additional two strain gauges are inserted in the Wheatstone bridge circuit, this is now a “full Wheatstone” force sensor (see [Figure 12-4](#)). Note, however, the diagonally symmetrical orientations of the strain gauges, necessary to multiply their effect.

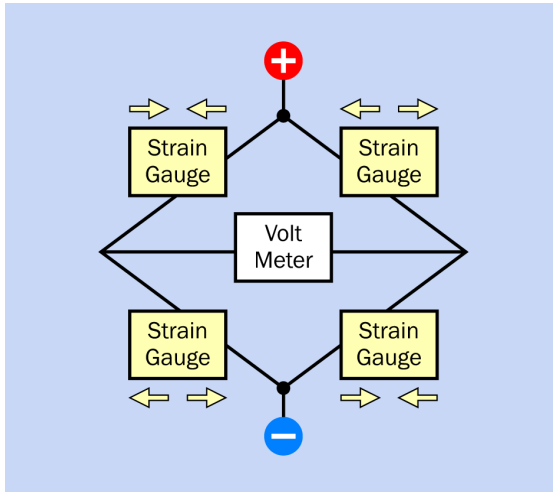


Figure 12-4 Two additional strain gauges create a “full Wheatstone” circuit.

A typical digital bathroom scale contains two half-Wheatstone force sensors, wired to create a full-Wheatstone configuration.

The force sensors shown in [Figure 12-5](#) are rated for up to 50kg each, and can thus weigh up to 100kg if they are combined in a scale. In this figure, one sensor has been turned over to show its underside. The strain gauges are hidden in each sensor where the steel sections overlap.

Wheatstone Bridge Errors

Where more than one strain gauge is used in a Wheatstone bridge, they should ideally have identical performance. Since this is impossible as a result of manufacturing tolerances, some error correction is built into devices using multiple strain gauges.

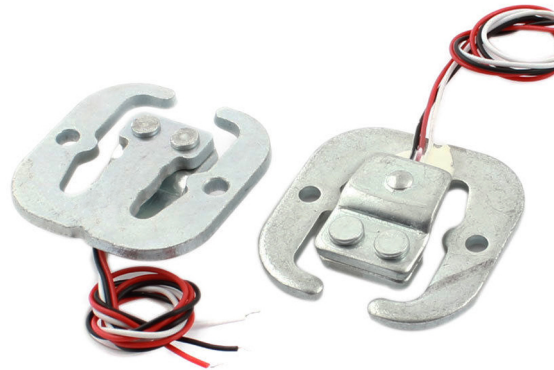


Figure 12-5 Two force sensors suitable for a bathroom scale, each containing a pair of strain gauges that will flex oppositely under load.

Strain-Gauge Amplification

The voltage output from a Wheatstone bridge circuit is given by the following formula, where V_{IN} is the supply voltage, V_{OUT} is the output, and $R1$ through $R4$ are the resistance values that were used in [Figure 12-2](#).

$$V_{OUT} = [(R3/(R3+R4)) - (R2/(R1+R2))] * V_{IN}$$

The good news is that the output when a strain gauge is used will vary linearly with the load applied. The bad news is that it will be very small.

To amplify it, an op-amp such as the AD620 is often recommended. Using external resistors, its amplification factor can be adjusted from 1:1 to 10,000:1. Alternatively a chip such as the HX711 by Avia Semiconductor contains a 10-bit analog-to-digital converter and is specifically designed for use in weighing scales. Its digital output uses a very simple serial format. Sparkfun sells a breakout board incorporating this chip. See [Figure 12-6](#).

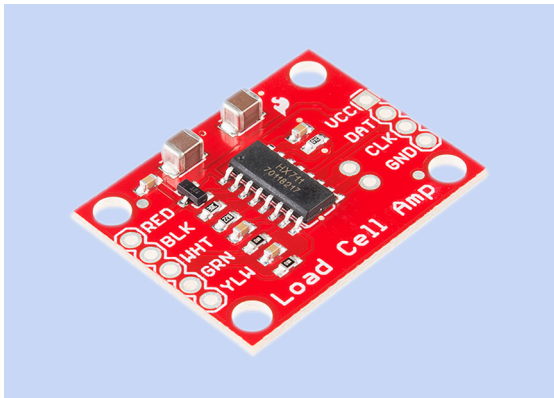


Figure 12-6 Sparkfun HX711 breakout board for the HX711 amplifier chip, specifically designed for a full-Wheatstone array of strain gauges in a force sensor.

Other Strain-Gauge Modules

Strain gauges are built into a variety of sensor modules. Figure 12-7 shows a selection, all of which are available from Sparkfun. Many more can be found online.

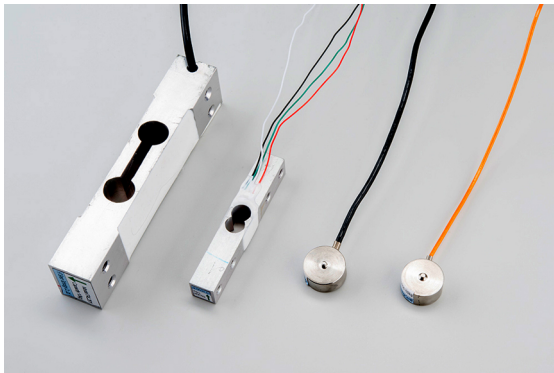


Figure 12-7 A selection of load cells containing strain gauges, available from Sparkfun.

Plastic-Film Force Sensors

Plastic-film resistive sensors contain two layers of conductive ink, sealed between two layers of thin, transparent plastic film. The resistance between the ink layers diminishes when they are pressed together. It may vary from as little as 30K when fully loaded to more than 1M when unloaded.

Like a strain gauge, this sensor has no polarity and requires no power supply.

Examples of this type of sensor are shown in Figures 12-8 and 12-9. Other shapes and sizes are available. Manufacturers include Tekscan, whose product is named FlexiForce; Alpha Electronic, in Taiwan; and Interlink Electronics.



Figure 12-8 A FlexiForce A401 resistive film sensor made by Tekscan, Inc. Its sensing area is 25.4mm in diameter (1 inch) and is rated to measure up to 111N (25lbs).

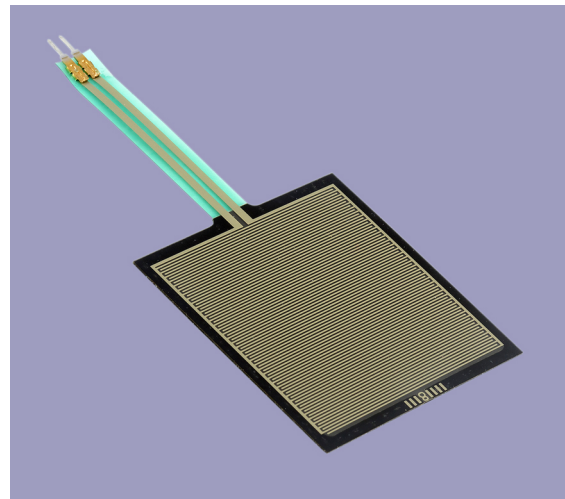


Figure 12-9 An Interlink FSR406 resistive film sensor. Its sensing area is slightly less than 40mm square, and is rated to measure up to 20N (4.5lbs).

Plastic-film sensors should not be confused with film-based piezoelectric vibration sensors, which are described in the **vibration** entry (see “Piezoelectric Strip”). Those sensors provide a transient output when they flex rapidly. The

resistive sensors provide a stable output in response to a steady load.

Deformative Force Sensors

A sheet of natural or silicone-based rubber can be impregnated with conductive particles. The conductivity of the sheet may not change significantly when it is compressed, but if it is separated from a metal plate by a mesh of thin nylon fibers, compression will result in greater conductivity by pushing the rubber into the gaps between the fibers. See [Figure 12-10](#).

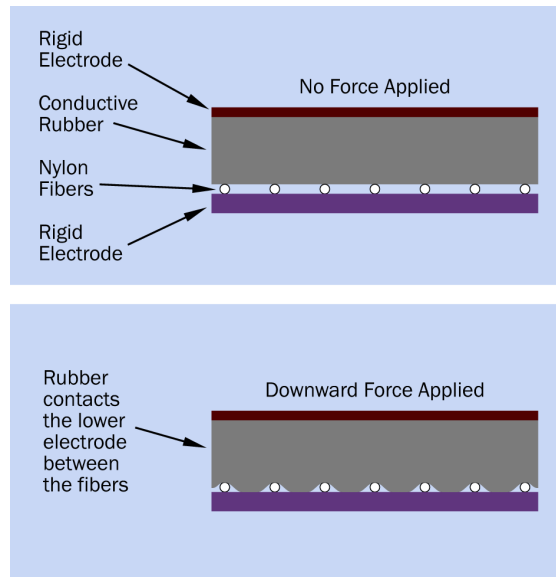


Figure 12-10 In the top image, a flexible conductive layer is separated from an underlying rigid electrode by nylon mesh. In the lower image, a load applied to the flexible layer has forced more of it into contact with the electrode, reducing the resistance between them.

Improvised Resistive Sensors

Polyethylene film impregnated with carbon particles is available under the brand name Velostat, owned by 3M. Although it was developed as an antistatic packaging material for semiconductors, it can be used to make a DIY force sensor. When the material is stretched, the embedded particles are more dispersed, and electrical resistance increases. When the material is compressed, its resistance is reduced.

Antistatic foam of the type used to package CMOS components can be used in the same way, although some types behave like memory foam, being slow to recover from pressure. The foam can be sandwiched between a pair of copper-plated circuit boards as electrodes.

How to Use It

Plastic-Film Resistive Force Sensors

The conductivity of a plastic-film sensor has an almost linear relationship with the force applied. In other words, if F is the force and I is the current:

$$I = k * F$$

where k is a constant determined by the characteristics of the materials used.

By Ohm's Law, $R = V / I$ where V is the voltage drop across a resistance of value R . By substitution, using $k * F$ instead of I :

$$R = V / (k * F)$$

Therefore, if a constant voltage is applied across the force sensor, the resistance of the sensor will be proportional to the reciprocal of the force (i.e., $1 / F$). These relationships are shown in [Figure 12-11](#).

For convenient measurement, it will be helpful if the resistance of the sensor can be converted to a voltage output that varies linearly with the force applied. To achieve this, a resistive force sensor of this type is customary amplified with an **op-amp** (see Volume 2). Using the schematic shown in [Figure 12-12](#), the amplification ratio, A , is found from this formula:

$$A = 1 + (R2 / R1)$$

where $R2$ is the potentiometer, and $R1$ is the resistive force sensor. Therefore, the output from the op-amp should have an approximately linear relationship with the force applied to the sensor.

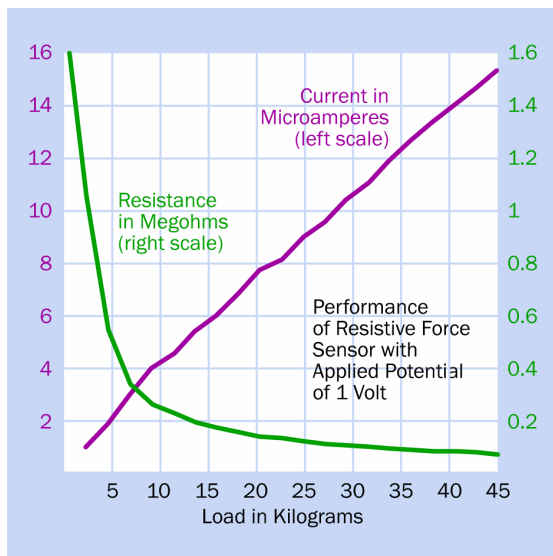


Figure 12-11 The relationship between force, current, and resistance in a flexible resistive sensor, assuming a constant voltage of 1V is applied across it (derived from FlexiForce Sensors User Manual).

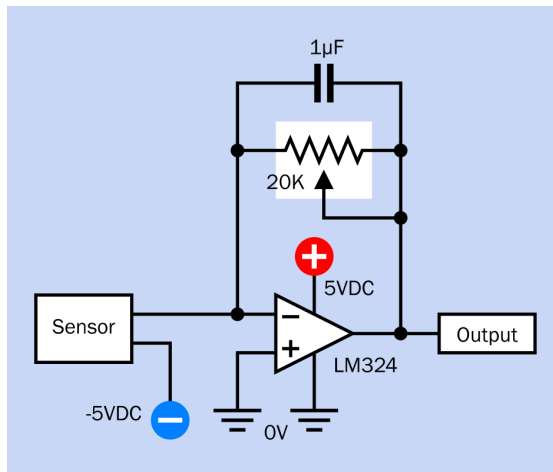


Figure 12-12 An amplification circuit with an output that is approximately linear with force applied to a resistive film sensor.

The capacitor in this circuit is added to suppress noise that the circuit might otherwise pick up.

An alternative is to wire a resistive sensor in series with a capacitor, and connect the other side of the capacitor to a comparator that has adjustable feedback. The resistance of the sen-

sor will determine how quickly the capacitor charges. However, because of the mathematics describing the charge rate, the output from the comparator would not be linear with the force on the sensor. Also, provision would be necessary to discharge the capacitor intermittently.

Values

Film-Based Force Sensors for User Input

A very light pressure with a fingertip could be around 50g. A more defined finger-press would be 250g, and a heavy push with a finger would be around 1kg.

The Interlink range of flexible force sensors requires a minimum pressure of 0.2N, or about 20g. Similarly, Alpha products range from a minimum of 10g to 30g.

These specifications suggest that film-based force sensors may be used for one-touch user input, but the no-load resistance will be at least 1M, and more than 10M in some instances. A small amount of pressure is unlikely to reduce this electrical resistance much below 500K. Using an op-amp or comparator to detect that difference and convert it into a reliable on-off output may be vulnerable to noise and power-supply disturbance.

Another consideration is that film-based sensors provide no tactile feedback. For these reasons, and because film-based products are described by their manufacturers as “force sensors” rather than as “touch sensors,” they are included in this entry rather than in [Chapter 13](#). That said, they may be considered as an option for one-touch user input where they are appropriate—for instance, in games where players are likely to slap or hit a sensor vigorously.

Specifications for Film-Based Force Sensors

Durability of film-based sensors is excellent, with manufacturers claiming that performance

is not degraded after 1 million applications of a 20kg load. Sensors have no polarity, will function using voltages from 1V to 15V in most cases, and have a response time of less than 5ms. They do not generate, and are not vulnerable to, electromagnetic interference.

Important attributes when evaluating film-type sensors are their limits of tolerance for applied force, and resistance values at each end of the range. Unfortunately these values are poorly documented for many sensors, or may be unspecified.

Maximum force may range from 20N to 440N, depending on the brand and model of sensor.

Electrical resistance is at least 1M when unloaded, and may be as high as 20M.

Accuracy ranges from plus-or-minus 2% to plus-or-minus 5% from one application of force to the next, depending on the model of sensor and the manufacturer. If force is not applied each time in exactly the same area of the sensor, results will vary. If a sensor is swapped with another sensor of the same type, sensor-to-sensor consistency will be uncertain. Therefore, film-based force sensors are not a good choice for applications where accuracy is important.

Sensors may be rated for 5% to 10% hysteresis.

The area of active detection may range from about 4mm wide (FlexiForce sensors) to more than 40mm wide (Interlink FSR-406).

Strain Gauges

Strain gauges are not sold as individual components by most electronics suppliers. They are sometimes available as surplus parts or from sites such as eBay, where specifications range from 100 ohms to 1K as the resistance when no load is applied.

A load sensor on which a strain gauge has been preinstalled is much easier to use, and will be plug-compatible with an appropriate amplifier chip as previously described.

What Can Go Wrong

Soldering Damage

The pins on plastic-film resistive sensors are embedded in thin plastic. Heat from a soldering iron can easily damage this plastic. Heat-sink alligator clips should be used while soldering, or the pins can be socketed instead of soldered.

Bad Load Distribution

Film-based sensors will not provide accurate readings if a load is imposed unevenly or inconsistently, or extends outside the detection area. A *puck* consisting of a small, rigid disc may be interposed between the source of the force and the sensor, to distribute the load within the maximum area. A puck may also be referred to as a *shim*.

Similarly, the sensor must be mounted on a flat, smooth surface, and if this is not available, a rigid plate should be interposed.

Water Damage

Although film-based sensors are enclosed in plastic, they are not waterproof. Immersion may cause the layers to delaminate.

Temperature Sensitivity

Because electrical resistance tends to vary with temperature, readings from resistive force sensors will vary with temperature.

Ambient temperatures of 70 degrees Celsius and above may damage a film-based sensor.

Leads Too Long

Although film-based sensors are supplied with a variety of lead lengths enclosed in the laminated layers of flexible plastic, the leads may be too long for a particular application. If they are trimmed, wires cannot be attached with solder, as it will melt the plastic. Conductive epoxy should be used.

single touch sensor

This entry only describes *capacitive* touch sensors. A *conductive* sensor, which uses the fingertip to complete a circuit between two exposed contacts, is not very common, and is not included here.

The type of touch sensor described in this entry requires no physical pressure for activation. It should not be confused with a resistive or piezoelectric **force** sensor that requires pressure. See Chapter 12.

An integrated circuit chip that processes a signal from a *touch pad* is often described as a *touch sensor*, even though it does not contain a sensing element. This entry describes it as a “touch sensor chip” to eliminate ambiguity, and refers to touch input elements as “touch pads.”

Touch pads that contain *tactile switches* or *membrane switches* are described in the entry discussing **switches** in Volume 1. All types of switches are described in that volume, with the exception of a *reed switch*, which is magnetically activated and is therefore categorized as a sensor.

Capacitive touch sensors are sometimes referred to as *capacitive proximity sensors*, because they sense the proximity of a human fingertip. In this Encyclopedia, and in most other sources, a **proximity** sensor measures distance, not touch. See Chapter 5.

A *capacitive displacement sensor* employs the same principle as a capacitive touch sensor, but is used to detect the position of an object, not for human input.

OTHER RELATED COMPONENTS

- **force** sensor (see Chapter 12)
- **touch screen** (see Chapter 14)

What It Does

A *touch pad* detects the presence of a human fingertip (or other part of the body) and signals an integrated circuit chip, which is very often termed a *touch sensor*, even though it does not contain a sensing element itself. The chip creates an output to signify that human touch has been recognized.

A keypad of the type found on microwave ovens may appear similar to an array of touch pads, but is more likely to contain *membrane switches* or *tactile switches*, which are described with other forms of switches in Volume 1. The type of touch pad described in this entry requires no physical force and contains no parts that move or flex when pressed.

A modern **touch screen** is usually a capacitive device and can be thought of as an array of touch pads. See [Chapter 14](#).

Applications

Capacitive touch sensors have become common as their cost has fallen relative to simpler components that respond to being pressed.

A touch sensor can be used to start or stop a process, or to power up or power down a device. Multiple sensors may be found wherever user input of a few alphanumeric characters is required. Because touch pads can be completely sealed, they are useful where hygiene is important.

Specific applications include the activation of a backlight in a handheld device, wake-up from standby, ear detection in a cellular telephone, control of medical devices, and activation of interior lighting in some automobiles.

The absence of moving parts or electrical contacts means that a touch pad is more reliable than any type of electromechanical switch. A disadvantage is that it provides no tactile feedback, and therefore will require a visual or audible confirmation when it responds to input. Lack of tactile feedback makes touch pads unsuitable for computer keyboards and other key-entry devices where rapid typing is required.

When the capacitive elements of a touch pad are transparent, it can be mounted in front of a screen.

Schematic Symbols

Either of the schematic symbols in [Figure 13-1](#) may sometimes be used to represent a touch sensor, but not on a consistent basis.

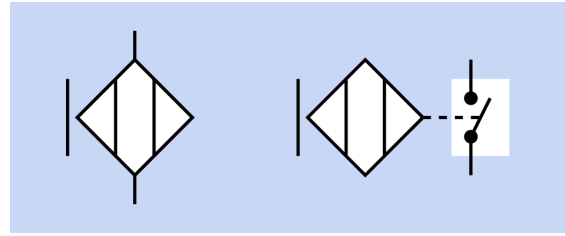


Figure 13-1 Two possible schematic symbols that may represent a touch sensor.

How It Works

The two plates in a capacitor are separated by an insulator known as the **dielectric**. Paper, plastic, glass, air, and other insulators can serve this purpose. Although no electrical connection exists between the plates, AC passes through the dielectric as a field effect.

Electrical capacitance exists between any pair of electrical conductors. The human body has high electrical resistance but still is electrically conductive, and therefore has capacitance with other conductive objects.

A touch pad can function as one side of a capacitor, with a fingertip functioning as the other side. In this mode, AC can pass from the touch pad and through the human body to ground. The current is very small, but fluctuations can be detected by an appropriately designed integrated circuit chip, or by a microcontroller.

The precise characteristics of a dielectric will affect the performance of a capacitor to some extent, but will not prevent it from working. Therefore a touch pad can function even when it is shielded behind a protective layer of glass or plastic, as is often the case.

A touch sensor chip generates pulses of low-voltage AC and sends them to a touch pad. The chip detects any variation in the current through the touch pad, indicating the presence of a fingertip. Where an input occurs, the chip changes its output, which usually requires a microcontroller for processing.

How to Use It

Touch sensor chips are available in as many as 40 different formats and configurations. All of them are surface-mount. For breadboarding, an experimenter can use breakout boards that have sensor chips installed. [Figure 13-2](#) shows a product from Adafruit that is capable of addressing 12 touch pads. Its output is accessible from a microcontroller via the I2C protocol. For additional details about protocols such as I2C, see [Appendix A](#).

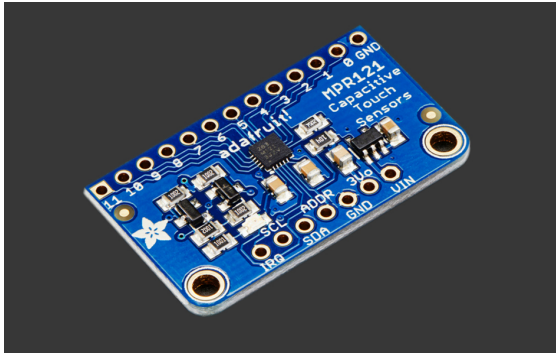


Figure 13-2 A capacitive touch sensor chip on a breakout board from Adafruit.

Similar breakout boards are available from Sparkfun, and from large online vendors such as <http://www.mouser.com> (where they are categorized as *development tools*).

While most touch sensor chips require a microcontroller, a few are available with the same number of output pins as touch-pad input pins, and each output pin will transition between logic-high and logic-low when an input on the corresponding pin is detected. Another breakout board from Adafruit, the AT42QT1070, uses this simple system.

A [library](#) exists for Arduino that enables two pins to sense touch on a piece of aluminum foil.

It can also work with [conductive ink or paint](#).

Obtaining Touch Pads

Sensor chips are widely available as components, and are very inexpensive. On the other

hand, touch pads are not common as components, probably because a touch pad is usually created as a pattern of copper traces etched onto a circuit board by a device manufacturer.

Touch pads from hobby-electronics sources usually include touch sensor chips. Sparkfun offers a 12-key keypad on this basis, and also a 9-key keypad designed as a “touch shield” for use with the Arduino. Both of the Sparkfun products are shown in [Figure 13-3](#). They include the same MPR121 touch sensor chip as the breakout board from Adafruit shown in [Figure 13-2](#), and require an I2C connection with a microcontroller.

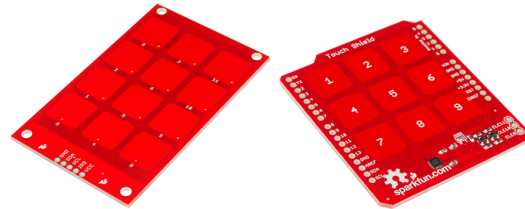


Figure 13-3 Two capacitive keypads from Sparkfun, the one on the right designed as an Arduino shield.

Because a capacitive touch pad is usually mounted inside an enclosure, the appearance of the bare touch pad as a component is unimportant. The outside of the enclosure can be printed with a design showing key outlines.

Individual Touch Pad

Adafruit sells the AT42QT1010 touch pad that emulates a momentary switch. Its output transitions from logic-low to logic-high when a finger presence is detected, and transitions back to logic-low when the finger is removed.

An alternative, shown in [Figure 13-4](#), toggles between a logic-high and logic-low output and latches in each state with each single key press.

Both of these keypads contain sensor chips to generate the output.

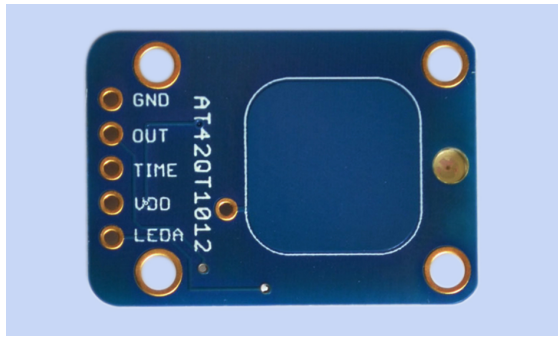


Figure 13-4 The output from the AT42QT1012 sensor chip on this touch pad from Adafruit toggles between logic-high and logic-low each time the pad is touched.

Wheels and Strips

A touch wheel uses a circular pattern of conductive traces, often referred to as *electrodes*, to receive finger input. A simple configuration is shown in Figure 13-5. The traces interlock without making contact with each other, so that moving the finger in a circular motion creates a capacitive input that rises and falls sequentially on each element. In the figure, three sections are used, each colored differently here for purposes of clarity. Other touch wheels may contain more sections.

Firmware that is designed to interact with wheel-shaped touch pads will generally assume that two electrodes are receiving input at one time. The firmware attempts to calculate the position and motion of the finger by assessing the relative capacitance of adjacent segments. Ideally the capacitance values should vary linearly and complementarily; that is, from a 50-50 value at a midpoint between two segments, the progression should change to 60-40, 70-30, and so on, as the finger moves around the wheel.

A *touch potentiometer* consists of multiple touch pads, usually arrayed in a straight line. This may be described as a *touch strip*. An example is made by GHI Electronics and sold by Robot Shop as their L12 Capacitive Touch Module, shown in Figure 13-6.

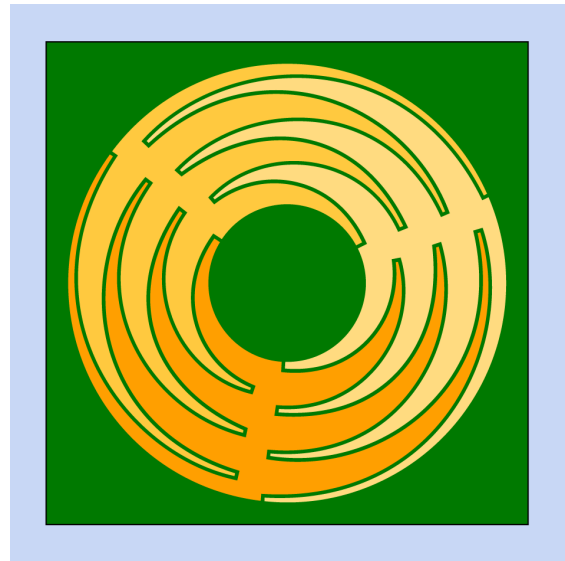


Figure 13-5 A touch wheel created by copper traces on a circuit board (green). Each section is colored differently, for purposes of clarity.



Figure 13-6 Multiple capacitive touch pads arrayed as a strip, made by GHI Electronics.

Design Considerations

A simple touch pad is often surrounded by additional copper that is grounded and may be described as a *shield* or a *guard*. The capacitance of a fingertip (spaced above the electrode by a layer of plastic or glass that functions as the dielectric) interferes with the field between the electrode and the shield.

The underside of the circuit board is often plated and grounded to protect the touch pad from electromagnetic interference. This, too, can be referred to as a shield or a guard. The ground plating can be in a hatched pattern to reduce its capacitance with the electrode above.

Layout of a circuit involving a single touch sensor chip can affect its performance significantly. A touch sensing device may have to be “tuned” to detect finger presence reliably.

A long trace between the touch sensor and an electrode will tend to pick up noise and will increase capacitance.

The distance between adjacent traces from multiple touch pads must be maximized to reduce capacitance between them. If output from a sensor chip uses the I2C or SPI digital protocol, any trace carrying that digital signal should be at least 4mm from input traces. If they cross, they should be at 90 degrees to each other.

Electrodes should not be shaped to resemble numbers or other characters printed above them. A single basic electrode should be circular.

What Can Go Wrong

Insensitive to Gloves

Gloves are a challenge for touch-sensor design, as they alter the dielectric and the distance between the electrode and the finger. Capacitive touch sensors may not work at all with some types of gloves. However, special gloves containing metallic threads are available.

Stylus Issues

A nonconductive stylus cannot activate a touch pad.

Conductive Ink

Ink that prints the shapes of touch pads on the exterior of a device should be nonconductive.

touch screen

14

The term **touch screen** is written as two words in this Encyclopedia. In many sources, the words are concatenated as “touchscreen.”

The two words are hyphenated here when they function as an adjective, but not otherwise. In manufacturers’ datasheets where the two words are used, they are not usually hyphenated.

OTHER RELATED COMPONENTS

- **single touch** sensor (see [Chapter 13](#))
- **force** sensor (see [Chapter 12](#))

What It Does

A touch screen is a video display with embedded touch sensing. The screen reports the position of the touch, and is used as a pointing device as an alternative to a mouse or trackpad. Some touch screens report pressure as well as position.

Touch screens are widely used in smartphones and tablets, and also in some laptop computers. Smaller, simpler touch screens may be found in office equipment such as photocopiers.

Schematic Symbol

No specific schematic symbol is used to represent a touch screen.

Variants

Early designs used infrared LEDs recessed into the edges of a frame around a screen. A matching photodiode picked up the focused beam from each LED. The presence of a fingertip was detected when it interrupted one or more of

the beams. This system was not capable of high resolution, but was adequate for detecting user input at predefined locations.

Most touch screens currently are either resistive or capacitive.

Resistive Sensing

A resistive touch screen consists of two transparent layers that can be installed over a separate video display.

Each of the layers has uniform electrical resistance. Pressure from a fingertip on the outer layer (which we can refer to as layer 1) forces it to make contact at a point with the inner layer (referred to here as layer 2).

Two vertical electrodes connect with layer 1 along its left and right edges. Two horizontal electrodes connect with layer 2 along its top and bottom edges. When voltage is applied between the vertical electrodes on layer 1, the layer acts as a horizontal voltage divider. The voltage at the point where the layer is being pressed is applied to layer 2, and can be read from either of the electrodes on layer 2, so long

as the metering has a much higher impedance than that of the layer. The voltage is decoded as a value for the horizontal position on layer 1.

An external switching device now repeats the procedure, except that it applies voltage to layer 2, and reads it from layer 1 to supply a value for the vertical position on layer 2. The sequence is illustrated by the top and bottom sections of [Figure 14-1](#).

Because only four connections are necessary, this is referred to as a *four-wire* resistive touch screen. Five-wire variants exist, but are less common, and are not included in this entry.

Advantages of a resistive screen include:

- Simplicity. Only four connections are necessary, and the layers of the screen do not have to be subdivided into separate conductors.
- Low cost, relative to capacitive touch screens.
- Will respond equally well if a user wears gloves or uses a stylus.

Disadvantages of a resistive screen include:

- Some resistive versions require a stylus input instead of finger pressure.
- Resistive screens only respond to a one-location input. Two-finger gestures are not supported.
- Contact bounce occurs when the flexible layer is pressed against the underlying layer, and voltage spikes may be associated with switching power to the screen. To address this issue, firmware in a microcontroller may have to take a median value from several rapid readings.
- The flexible membrane is vulnerable to damage from sharp objects.

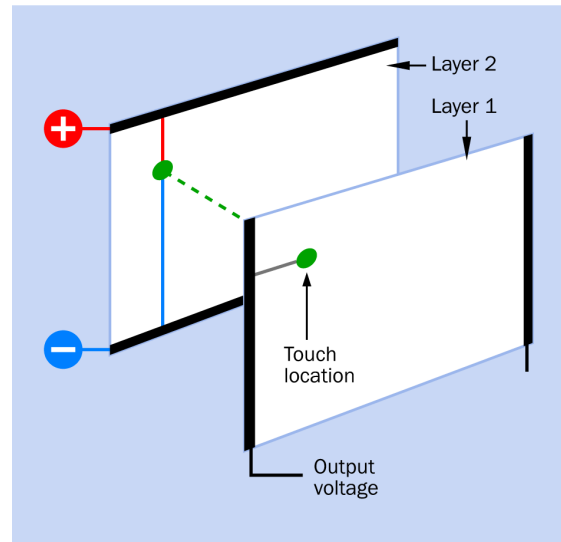
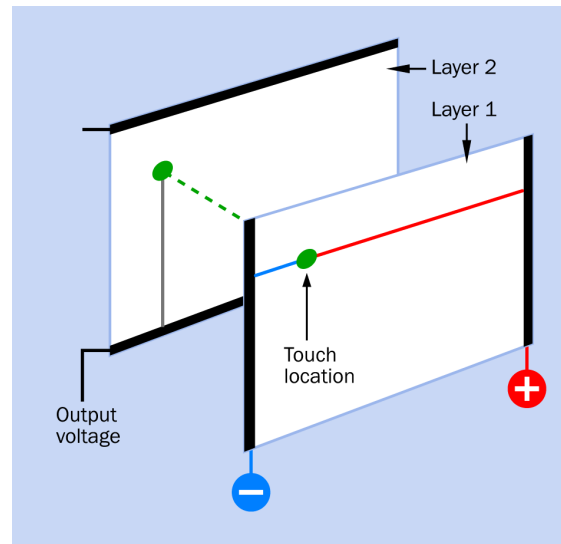


Figure 14-1 Sections of a resistive touch screen are shown displaced for clarity. In reality they would be separated by a very small gap, allowing one section to make contact with the other in response to finger pressure.

Capacitive Sensing

A capacitive touch screen can consist of an array of **single touch** sensors printed onto a glass panel as vertical and horizontal lines of transparent conductive ink.

Alternatively, a small capacitive screen can measure the tiny amount of current drained by

a fingertip from four sources located at corners of the screen.

For more information about capacitive touch sensing, see the entry on touch sensors in [Chapter 13](#).

Screens Available as Components

A wide variety of screens can be found in diagonal sizes ranging from 2 inches upward. Onboard electronics can be suitable for connection with a microcontroller using the I2C and SPI protocols, or USB. Different screen resolutions are available.

An example of a 3.5-inch touch screen with 320x200 resolution, mounted on a breakout board that can be used with a breadboard and an Arduino, is shown in [Figure 14-2](#).

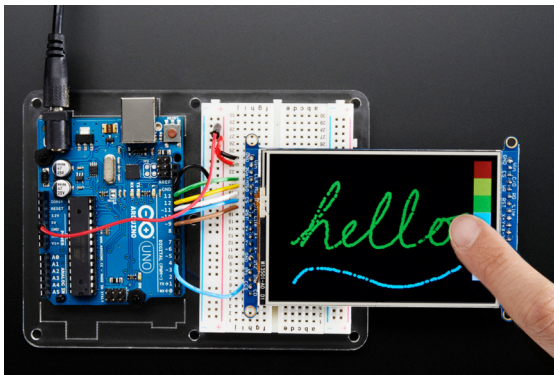


Figure 14-2 An Arduino-compatible touch screen mounted on a breakout board, available from Adafruit.

A 7-inch resistive touch screen that can be mounted on a separate video display is shown in [Figure 14-3](#). It can be used with the STMPE610 controller chip, which converts resistive screen values into digital coordinates and can be accessed by a microcontroller over both SPI and I2C. This surface-mount chip is available on a breakout board.

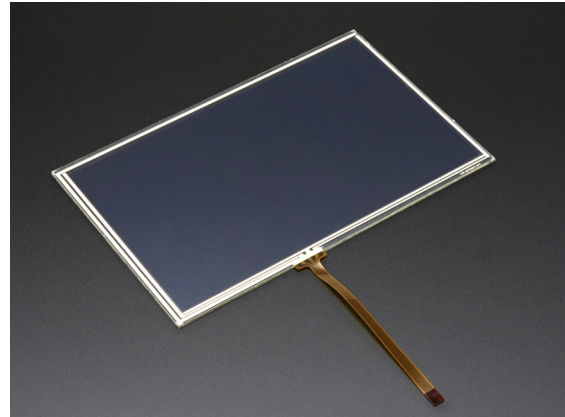


Figure 14-3 This resistive touch screen is intended for use as a layer applied to a 7-inch video display. The screen is available from Adafruit.

When choosing a touch screen as a component for a DIY project, the availability of microcontroller code libraries for reading and refreshing the display is an important consideration.

liquid level sensor

15

Level indicators that contain no electronic components are not included in this entry.

Specialized industrial level-sensing equipment is generally outside the scope of this Encyclopedia. This entry discusses small-scale, lower-cost sensors.

Liquid volume can be assessed by measuring liquid pressure at the bottom of a reservoir. Sensors for this purpose are discussed in the entry describing **gas/liquid pressure** sensors. See Chapter 17.

OTHER RELATED COMPONENTS

- **liquid flow rate** sensor (see [Chapter 16](#))
- **gas/liquid pressure** sensor (see [Chapter 17](#))

What It Does

Measuring the volume of liquid in a storage vessel or reservoir is such a fundamental task, countless methods have been devised, of which only the simplest and most common will be discussed here.

A liquid level sensor can have a binary output, meaning that it signals when the volume rises above or falls below a level that can be preset or reset. Often the sensor will be connected to a pump or valve that maintains a relatively constant volume in a container.

Alternatively a sensor can indicate the actual stored volume, either with an analog output or in digital increments.

Schematic Symbols

Three variants of a schematic symbol for a simple liquid level sensor are shown in [Figure 15-1](#). They are not always used, however, and a sensor may be shown simply as an annotated switch.

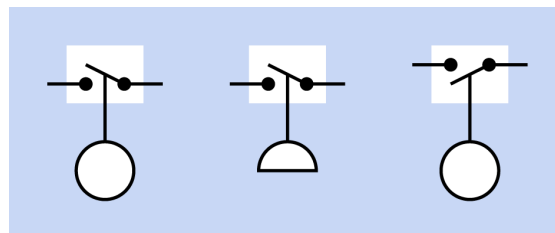


Figure 15-1 Three variants of a schematic symbol to represent a liquid level sensor. The rightmost symbol indicates that a rising level closes, rather than opens, a switch.

Applications

The fuel gauge in a vehicle is one of the most commonly encountered applications of a liquid level sensor. The water tank in a recreational vehicle or boat may use similar electronics. In industry, the choice of a sensor will be influenced by the type of liquid that is being stored, the desired accuracy, the temperature range, and whether the storage tank is sealed or open to atmospheric pressure.

How It Works

Desirable attributes of any liquid-level sensor include resistance to vibration, some damping to average out fluctuations caused by turbulence or sloshing in the liquid, resistance to chemical reactions with the liquid, and few moving parts that may require maintenance if the sensor is inside a sealed tank. Desirable attributes of an analog float sensor include a linear response, and some hysteresis if the application requires it.

This entry compares a variety of sensing strategies.

Binary-Output Float Sensor

The term “binary output” is used here to describe an output that only has two states (on and off, or logic-high and logic-low). The simplest type of liquid-level sensor with a binary output consists of a donut-shaped float that contains a permanent magnet and is free to slide vertically up and down a sealed tube containing a reed switch. The tube is supported on a bracket that can be mounted on the wall or the lid of the vessel containing the liquid.

The tube and float must be nonmagnetic, and the float must have a significantly lower specific gravity than the liquid that is being used (“significantly” because the float requires sufficient buoyancy to carry the weight of the magnet and overcome any friction between itself and the tube). A diagram illustrating this configuration is shown in [Figure 15-2](#).

To change the level setting of the sensor, the bracket may be mounted on a screw thread to adjust its vertical position.

The reed switch can be normally open or normally closed, as needed to respond to a rising or falling liquid level. For basic information about reed switches, see [“Reed Switch”](#). Additional, detailed information about reed switches is included in the book [Make: More Electronics](#).

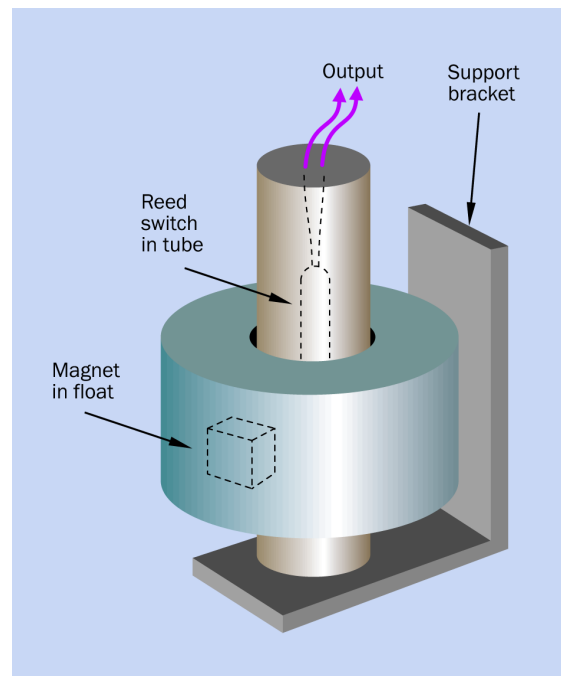


Figure 15-2 The principal parts of a basic binary-output float sensor.

For increased reliability, a Hall-effect sensor could be substituted for a reed switch. See [“Hall-Effect Sensor”](#) for general information about Hall-effect sensors.

Another binary-output float sensor is shown in [Figure 15-3](#). This is a sealed plastic capsule containing a snap-action switch and a steel ball. The cable is attached to the underside of the top of a tank, and the capsule dangles into liquid in the tank. A separate weight (not shown) has a hole in the middle, and is threaded over the wire. The weight keeps the wire approximately in a vertical position as it dangles into the tank.

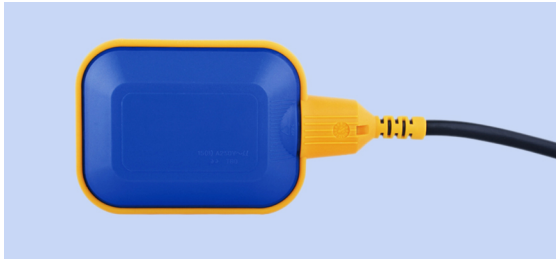


Figure 15-3 An air-filled float that switches external power depending on its orientation.

Figure 15-4 shows the components inside the float. When the liquid level in the tank falls, the float adopts the position shown on the left in the figure. The ball drops against a lever that closes a snap-action switch, which starts an external pump to replenish the tank. As the liquid level rises, the buoyancy of the air-filled float changes its orientation to that shown on the right in Figure 15-4. The ball drops and the switch opens, stopping the pump.

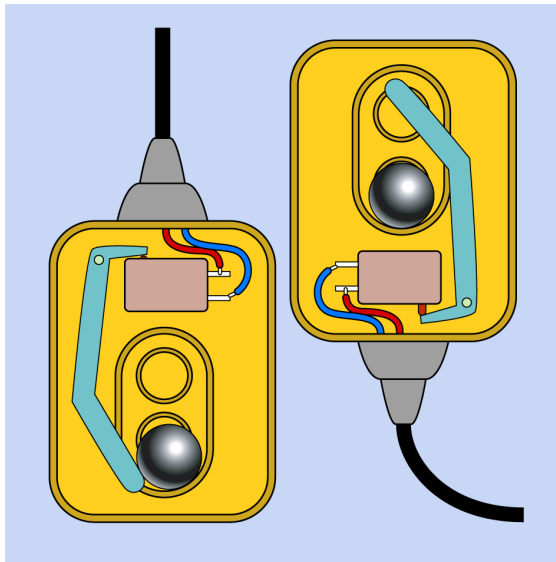


Figure 15-4 Internal components of the float shown in the previous figure.

Two circular indents on the inner surface of the plastic capsule prevent the ball from rolling erratically if there is turbulence in the liquid. They also provide some hysteresis.

Analog-Output Float Sensor

The simplest type of liquid-level sensor with an analog output consists of a float on an arm attached to a potentiometer, as shown in Figure 15-5. This very basic design was used in fuel tanks in vehicles for many decades. Disadvantages include a nonlinear response and the limited life expectancy characteristic of potentiometers. Some compensation for the nonlinear response can be made by using an analog fuel gauge with a nonlinear scale.

For more information about potentiometers, see “Arc-Segment Rotary Potentiometer”.

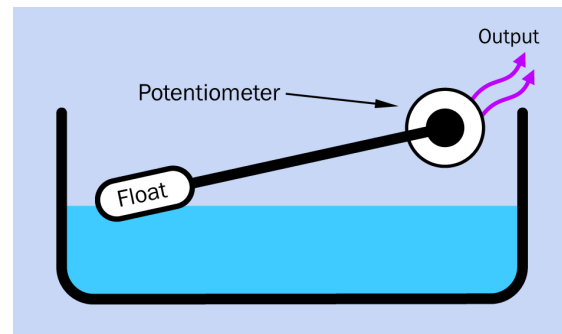


Figure 15-5 A basic float sensor with an analog output.

Incremental-Output Float Sensor

A schematic diagram for a simple float sensor with incremental output is shown in Figure 15-6. A magnet embedded in a donut-shaped float, similar to that shown in Figure 15-2, interacts with a sequence of reed switches installed in the central tube. The switches are spaced at equal intervals and apply power between resistors of equal value wired in series. This system has been used in motorcycle and automobile fuel tanks, where the switches may be enclosed in a (nonmagnetic) stainless-steel tube. The accuracy is limited by the number of reed switches.

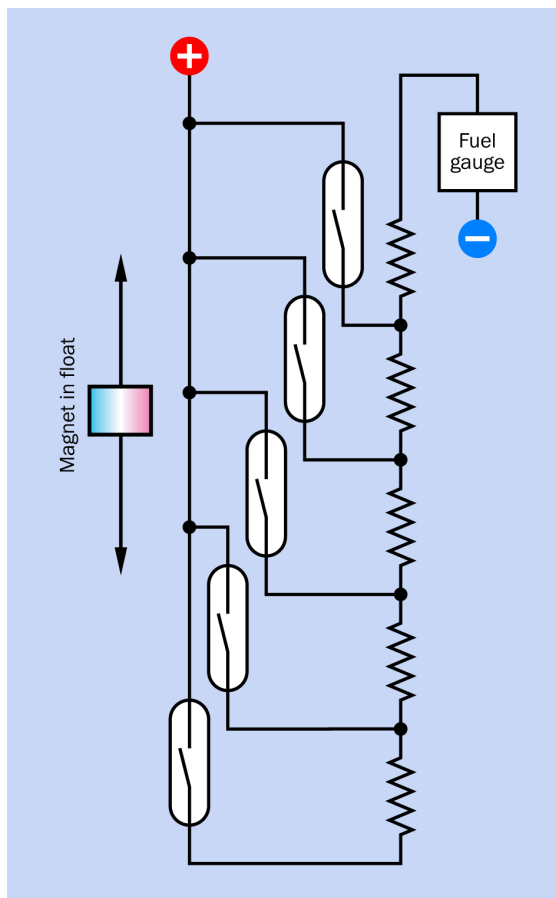


Figure 15-6 A float sensor with incremental output.

Displacement Level Sensors

If a heavy object, described as the *displacer*, is suspended in liquid, the effective weight of the object diminishes as the liquid rises up around it. This occurs because according to Archimedes' Principle, the upward buoyant force is equal to the weight of liquid that the object displaces. The displacer is suspended from a load sensor that measures its weight. Analog output from the sensor will be approximately linear with liquid level.

For a simplified diagram showing a displacement level sensor using this concept, see [Figure 15-7](#).

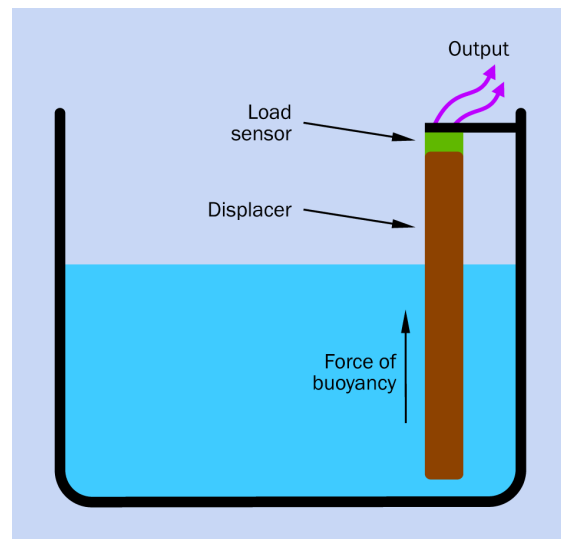


Figure 15-7 A displacement sensor. The displacer is heavier than the liquid around it, but its effective weight diminishes as the liquid rises.

For more information on load sensors, see [Chapter 12](#).

Ultrasonic Level Sensors

An ultrasonic sensor can be used to measure the level of liquid in a reservoir, as shown in [Figure 15-8](#). For more information about this type of sensor, see the entry discussing proximity sensors in [Chapter 5](#). A disadvantage of using ultrasound for liquid level sensing is that the speed of sound will be affected by any vapor given off by a volatile liquid.

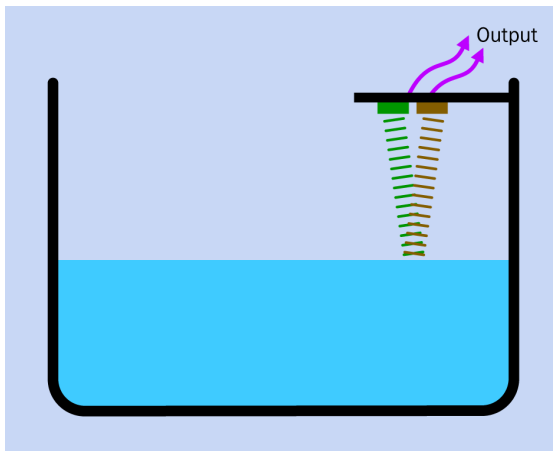


Figure 15-8 An ultrasonic proximity sensor can measure the level of liquid in a reservoir.

Reservoir Weight

The weight of a reservoir can be measured to assess the volume of liquid in it. This can be done by mounting the reservoir on load sensors. However, pipes leading to and from the reservoir must be designed so that they do not add or subtract any significant weight. [Figure 15-9](#) suggests an arrangement, although the outflow will still change the weight to some extent depending on the amount of suction that is applied. For more information on load sensors, see [Chapter 12](#).

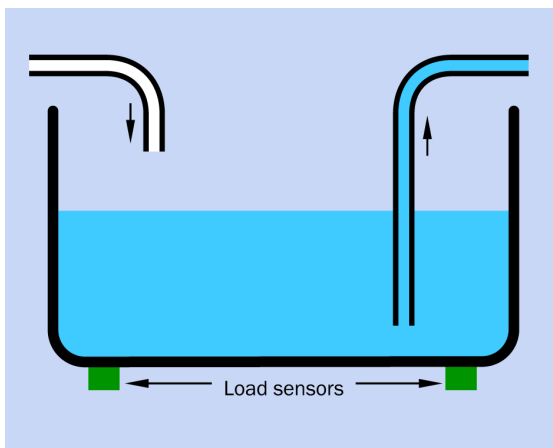


Figure 15-9 Using load sensors to assess the weight of liquid in a reservoir requires that the weight of plumbing should not be imposed on the structure of the reservoir.

Pressure Sensing

A differential pressure sensor can be added to a pipe near the bottom of a reservoir. The sensor measures the difference between the liquid pressure and ambient air pressure. See [Figure 15-10](#).

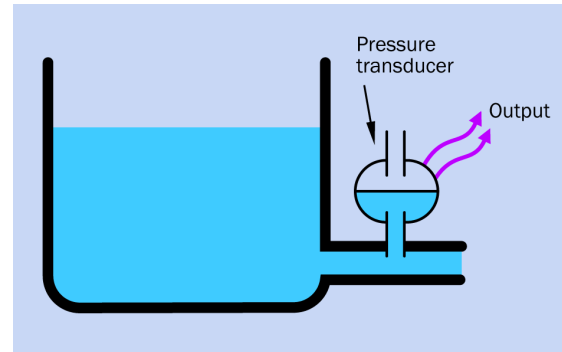


Figure 15-10 A pressure sensor can assess the volume of liquid in a reservoir.

This arrangement assumes that the reservoir is vented so that air pressure above the liquid level is equal to air pressure at the sensor. If the reservoir is not vented, a pipe must connect the reference port on the sensor with the space above the liquid.

The reservoir must have straight, vertical sides for the pressure to be directly proportional to the liquid volume.

Liquid volume can also be assessed by measuring the pressure inside a container, near the bottom. A submersible pressure sensor can be used, typically consisting of a watertight capsule fitted with a diaphragm that connects with an internal strain gauge. The sensor is lowered on a cable that also contains an air line. Because pressure in a liquid is affected by atmospheric pressure above the surface of the liquid, the sensor requires an air line so that its measurements are relative to the outside air.

Submersible pressure sensors are useful where access is limited—for example, when measuring fluctuations in an open-air municipal water reservoir.

What Can Go Wrong

Turbulence

Surface turbulence consisting of ripples, waves, or sloshing of liquid will tend to occur when a reservoir is refilled rapidly or is subjected to lateral movements as a result of being mounted in a moving vehicle. To minimize output fluctuations, some *damping* is desirable.

Baffles consisting of perforated plates inside the reservoir are a common strategy, as shown in [Figure 15-11](#). In the upper section of the figure, lateral acceleration causes submersion of a float sensor in the reservoir. In the lower section, perforated baffles minimize the problem.

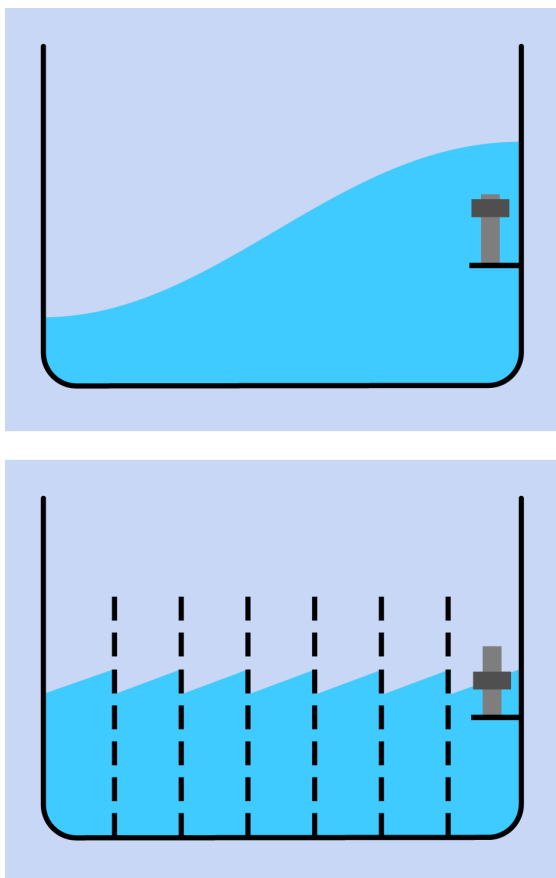


Figure 15-11 Insertion of perforated baffles in a reservoir minimizes the sloshing that otherwise tends to occur when the reservoir is subject to lateral motion.

Sensors that measure the weight or pressure of a liquid are less susceptible to turbulence. In a displacement sensor, the weight of the displacer provides a damping effect.

Tilting

All level sensors will tend to give inaccurate readings when a reservoir is tilted. A float sensor will be affected less if it is mounted centrally in a reservoir, because the reservoir will tilt around the sensor, as shown in [Figure 15-12](#).

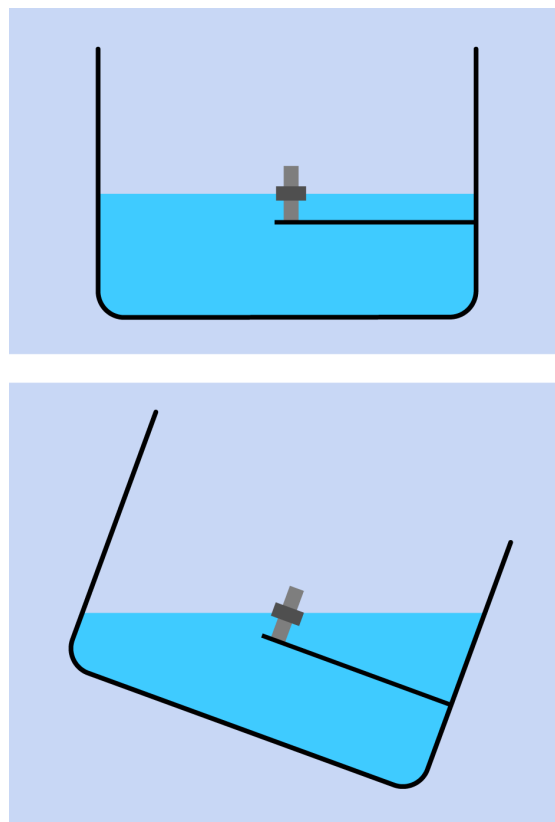


Figure 15-12 If a float sensor is mounted centrally in a reservoir, it will be significantly less affected if the reservoir tilts.

liquid flow rate sensor

18

Flow rate sensors that contain no electronic components are not included in this entry.

Some methods of liquid flow sensing can also be applied to gases, but sensors are usually designed for one application or the other. Therefore, gas flow sensors have their own entry. See Chapter 19.

Many liquid flow rate sensors are large devices designed for industrial applications. This entry focuses on lower-cost solid-state sensors.

OTHER RELATED COMPONENTS

- **liquid level** sensor (see [Chapter 15](#))
- **gas/liquid pressure** sensor (see [Chapter 17](#))
- **gas flow rate** sensor (see [Chapter 19](#))

What It Does

A liquid flow rate sensor measures the rate at which liquid flows past or through the device. A water meter is an example of a flow rate sensor.

A sensor may have a binary output, meaning that it signals when flow stops or starts, or if its rate rises above or falls below a level that can be preset or reset. However, most flow meters have an analog output that varies with the volume per unit of time.

Measuring the flow rate of a liquid can be challenging if the viscosity is very high, the liquid is chemically reactive, or the rate is very low. Such factors may require specialized equipment that is outside the scope of this Encyclopedia.

This entry compares the most popular sensing strategies.

Schematic Symbols

Many specialized symbols are used in flow diagrams to represent pumps, valves, and sensors. Typically they involve a single letter or an X in a circle. These symbols are not generally found in electronic schematics, and therefore they are not included here.

Paddlewheel Liquid Flow Rate Sensors

The simplest and most common liquid flow rate sensor uses a *paddlewheel*, also referred to as a *rotor*, that is mounted with its axis of rotation at 90 degrees to the direction of liquid flow. An example is the Koolance INS-FM16 shown in [Figure 16-1](#). This sensor is intended for use in an aftermarket cooling system for the CPU in an overclocked desktop computer, but can be used in any system where the rate of flow ranges from 0.5 to 15 liters per minute. The paddlewheel has a pair of magnets mounted in

it, activating a [reed switch](#) that is mounted in a sealed enclosure beneath the wheel. (For more information regarding reed switches, see “[Reed Switch](#)” in the entry describing **object presence** sensors.)

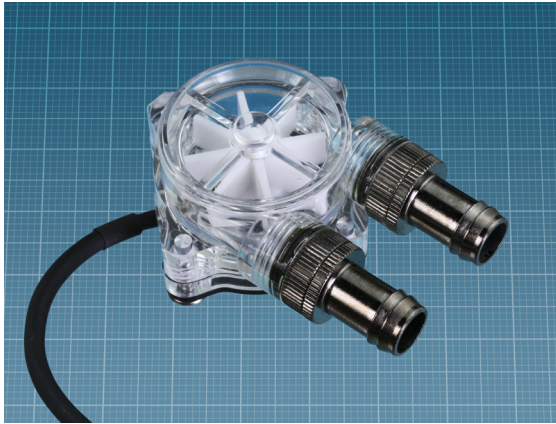


Figure 16-1 A low-cost, simple paddlewheel sensor designed for rates of 0.5 to 15 liters per minute. The background grid is in millimeters.

While the reed switch in the Koolance flow sensor inevitably suffers from contact bounce, it has the advantage of simplicity and can be used in conjunction with appropriate hardware or microcontroller code to debounce the pulse stream.

As in any device with a rotating part, friction and wear will afflict a paddlewheel sensor, especially because the bearings are often in the chamber through which the liquid passes. This eliminates the possibility for roller bearings. Typically a plain bearing is used, consisting of a pin that engages in a hole in the casing. Friction wears the bearing surface, creating a larger gap, which allows the rotor to vibrate or bounce instead of spinning smoothly. Flow resistance increases and accuracy is degraded.

In modern designs, the mass of the rotor is minimized to reduce the friction. Also, if the shaft is horizontal (but still at 90 degrees to the direction of the flow), buoyancy of the rotor in the liquid can reduce friction still further, as suggested in [Figure 16-2](#). Ideally, the density of

the rotor and the density of the liquid will be the same.

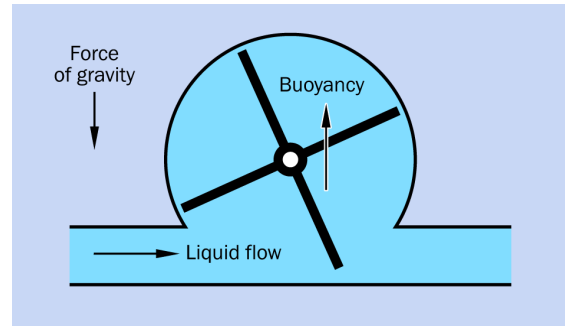


Figure 16-2 In this configuration, friction on the rotor bearings is mitigated by taking advantage of the buoyancy of a low-density rotor in the liquid that passes through.

The U-shaped flow path in the Koolance sensor in [Figure 16-1](#) maximizes the responsiveness of the rotor, but an inline path is more common. An example is shown in [Figure 16-3](#).



Figure 16-3 An inline flow sensor rated for 3 to 6 liters per minute.

Turbine Flow Rate Sensors

In a turbine-type sensor, two or more spiral blades are attached to a hub that rotates around an axis in line with the liquid flow, as shown in [Figure 16-4](#). A magnet in each spiral blade triggers a reed switch or Hall-effect sensor mounted in a bracket that suspends the turbine from the interior walls of the tube.

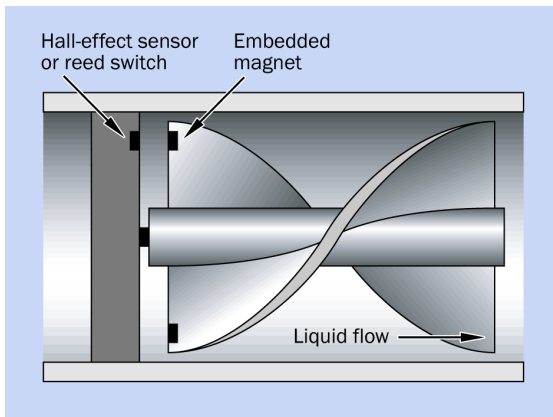


Figure 16-4 Simplified view of a turbine flow rate sensor mounted inside a tube.

The support bracket often consists of four struts, which impose resistance to liquid flow. The bearings suffer from the same kinds of problems as the bearings in a paddlewheel sensor, and must withstand additional load resulting from the inline force exerted by flow. Overall, while the turbine type of sensor is popular in laboratory equipment, it has disadvantages that are not shared by the bulkier paddlewheel type.

Limitations of Paddlewheels and Turbines

Both the paddlewheel and turbine types of sensors require a minimum flow to overcome the friction in their bearings. Below this minimum, liquid will find its way around the rotor without turning it. Even when the flow exceeds the minimum, response of the rotor is likely to be non-linear as a result of turbulence and other factors.

Above a limit stated by the manufacturer, turbulence increases to the point where output from the sensor is no longer meaningful. Wear on the bearing will also increase with flow rate.

A significant problem for these types of sensors is that they do not respond well to sudden variations in flow. The paddlewheel, in particular, has inertia as a function of the diameter of the rotor, and will take some time to spin up in

response to an increase in flow. Conversely, when the flow diminishes, the paddlewheel will tend to overrun.

The viscosity of the liquid passing through a paddlewheel or turbine sensor will have a very significant effect on its performance.

Thermal Mass Liquid Flow Rate Sensor

The thermal-mass system is commonly used when volumes are extremely low. The system is illustrated in Figure 16-5. A tube containing liquid is fabricated from a heat-conductive metal such as aluminum. It is enclosed in a larger tube, and the gap between them is filled with thermal insulation. A temperature sensor such as a thermistor measures the temperature of liquid entering the system. A second sensor, combined with a small resistive heater in the form of a coil around the tube, is placed downstream. Liquid passing through the tube will tend to remove heat more effectively at higher flow rates, and the difference in temperature between the two sensors is a logarithmic function of the flow rate.

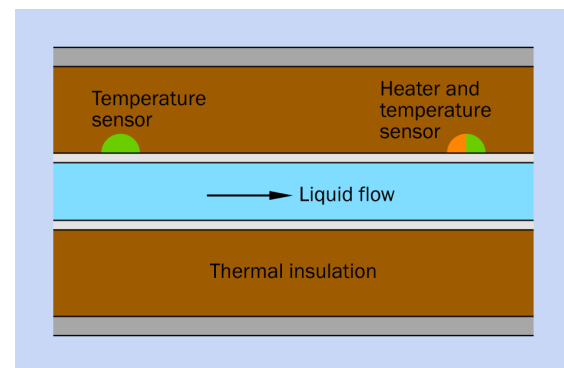


Figure 16-5 In this type of low-flow sensor, the temperature differential between the two sensors is a logarithmic function of the flow rate.

Variants of this system use slightly different tube configurations and sensor placement, but the principle is the same. Its advantages include the lack of any moving parts, and the

avoidance of any probes intruding into the liquid, which is desirable in biochemical and medical applications.

The same principle is applied in many **gas flow rate** sensors. See “[Mass Flow Rate Sensing](#)”.

Sliding Sleeve Liquid Flow Switch

This sensor is used in some domestic systems, where flow-activated water heating is required. A vertical section of brass (nonmagnetic) water pipe contains a sliding inner sleeve that incorporates a magnet. An external reed switch is activated when the sleeve is moved by water flowing through it. When the flow stops, the sleeve is returned to its rest position by the force of gravity.

Sliding Plunger Liquid Flow Switch

Figure 16-6 shows an exploded view of a similar device, using a plastic plunger that slides inside a nylon plumbing fixture designed for 3/4-inch pipe. The plunger contains a magnet and is restrained by a compression spring and a circular perforated plate. When water flow is sufficient to overcome the resistance of the spring, the plunger slides far enough for the magnet to activate a reed switch sealed into the external housing.

Ultrasonic Liquid Flow Rate Sensor

This type of sensor passes ultrasound through a liquid in a pipe. The speed of sound through the liquid is affected by the flow rate, and external electronics translate this lag time into a value for volume-per-minute. The system adjusts for variations in temperature that also affect the speed of sound.

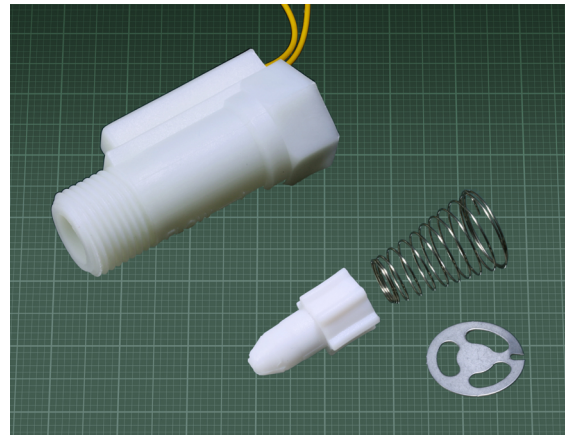


Figure 16-6 Parts of a flow switch. The small plunger is inserted into the pipe and retained with the perforated plate and compression spring.

Various configurations are available, some allowing ultrasound sources and detectors to be clamped to the outside of a pipe, as shown in Figure 16-7. To eliminate other variables, one ultrasound pulse is transmitted in the same direction as the flow, followed by another pulse contrary to the flow, and the difference between the two transmission times is used as an indicator of the flow rate.

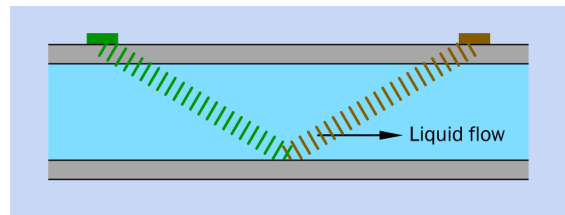


Figure 16-7 Some ultrasound flow sensors are designed to be clamped externally to a pipe.

Magnetic Liquid Flow Sensor

A magnetic field is induced in a metal pipe by a coil generating its field perpendicular to flow. The inside of the pipe is lined with nonconductive material in which two electrodes are mounted. Because water containing ions is conductive, the flow of water through the magnetic field induces a small potential difference

between the electrodes. This voltage can be used as an indication of the flow rate.

To eliminate external factors that would also affect the potential difference, the polarity of current through the coil around the pipe alternates rapidly. The induced field remains the same regardless of the direction of the current.

A magnetic flow sensor should not be confused with a magnetic flow [switch](#). Various types of switches are made, including large, heavy-duty devices where flow moves a magnet that triggers a shutoff valve. This type of industrial device is outside the scope of the Encyclopedia.

Differential Pressure Liquid Flow Meter

In this system, a pipe contains a perforated plate or some similar constrictor that partially obstructs the flow of liquid. Pressure is measured by a pair of pressure transducers placed before and after the constrictor. The pressure difference is an indicator of flow, because it increases as the flow rate increases.

This system was developed originally for large industrial applications but has been miniaturized and etched into silicon to measure very small flow rates. The Omron D6F-PH is an example, measuring less than 3cm square. It contains digital correction to enable a close-to-linear output. Because of its small size, it can be used only for slow flow rates, or as a bypass sensor. The general concept of a bypass sensor is illustrated in [Figure 19-7](#), in the entry discussing gas flow sensors.

What Can Go Wrong

Vulnerability to Dirt and Corrosive Materials

MEMS liquid flow rate sensors containing very delicate, very small sensing elements are very vulnerable to contamination with dirt. Liquids should be filtered to minimize this risk. A manufacturer's datasheet should provide information about the use of corrosive or chemically active liquids.

gas/liquid pressure sensor

17

Pressure measurement devices that do not contain any electronics, such as a nondigital tire-pressure gauge or a mercury manometer, are outside the scope of this Encyclopedia.

Many pressure sensing methods can be used with both gases and liquids. To avoid duplication, gas pressure sensors and liquid pressure sensors do not have separate entries. Both types are described here.

This entry deals almost exclusively with MEMS components. It does not include pressure measurement devices sold as industrial products.

Some manufacturers and vendors use the term *pressure sensor* to describe a component that measures mechanical load or force. In this Encyclopedia, the term only describes components that measure liquid or gas pressure. Mechanical load cells and force sensors will be found in the section on **force** sensors. See Chapter 12.

OTHER RELATED COMPONENTS

- **liquid level** sensor (see [Chapter 15](#))
- **liquid flow rate** sensor (see [Chapter 16](#))
- **gas flow rate** sensor (see [Chapter 19](#))

What It Does

A pressure sensor measures the force exerted by a gas or liquid, often in a container or pipe.

Static pressure is measured under conditions that change slowly or not at all. *Dynamic pressure* is subject to fluctuations. Pressure sensors tend to be designed for one condition or the other.

Schematic Symbols

Many specialized symbols are used in flow diagrams to represent pumps, valves, and sensors, including pressure sensors. Typically they involve a single letter or an X in a circle. These symbols are not generally found in electronic

schematics, and therefore they are not included here.

Applications

Barometric sensors are found in barometers and weather stations. Altimeters are really a specialized form of barometric sensor, used in airborne vehicles. Gas pressure sensors have many industrial applications, and are used to monitor tire inflation in vehicles and the output from air compressors. They may also measure liquid pressure indirectly, as in a blood-pressure cuff.

Liquid pressure sensors are widely used to measure oil pressure in automobile engines and hydraulic braking systems. They have med-

ical applications, and monitor water pressure in municipal systems and pumped supplies.

Design Considerations

As a result of random molecular movement, gases will tend to disperse to fill a container. After a gas reaches equilibrium, the pressure will be almost equal in all directions, affected only slightly by gravity. Where a gas is in a sealed, rigid container, pressure will vary linearly with temperature.

A liquid will tend to accumulate at the bottom of any container under the force of gravity. Liquid pressure in a container will be highest at the bottom, because of the weight of liquid above it. However, because almost all liquids are not easily compressible, they transmit force from any point in a container to any other point, including the sides of a container.

Units

Pressure is measured as force per unit area, which can be expressed in a confusing variety of units.

In the United States, gas and liquid pressure are still often expressed in pounds per square inch, abbreviated as PSI, or more often as psi, or sometimes as lb/in².

In standard international (SI) units, 1 bar of pressure is approximately equal to atmospheric pressure at sea level. Millibars are popular among meteorologists, 1 millibar being 1/1000 of a bar, equivalent to 100 pascals, where 1 pascal = 1 newton per square meter.

A bar is equivalent to 14.504 psi.

Blood pressure is measured in millimeters of mercury, because mercury manometers were used for this purpose originally. Atmospheric pressure may also be measured in millimeters of mercury, because the earliest barometers used a tube containing mercury. In the United States, some sources still refer to inches of mercury.

How It Works

Pressure sensing usually entails three stages.

1. A *sensing element* translates pressure into the mechanical displacement of a flexible part.
2. A *transducer* converts mechanical displacement into an electrical effect, either modifying resistance or creating a small voltage or current.
3. Electronics are used for *signal conditioning*. This may entail modifying a nonlinear signal, or may convert an analog output to a digital output.

As pressure sensors are increasingly taking the form of MEMS devices, all three stages may be combined in one silicon chip.

Basic Sensing Elements

Figure 17-1 shows four types of sensing elements that are used, or have been used, to convert pressure into mechanical motion. In each case, a green arrow shows where gas or liquid is introduced under pressure. 1: A Bourdon tube flexes under pressure, increasing its radius. The tube is hollow, open at one end and sealed at the other. 2: A coiled Bourdon tube uncoils partially under pressure, causing the top end to rotate. 3: A simple flat diaphragm. 4: A ribbed diaphragm.

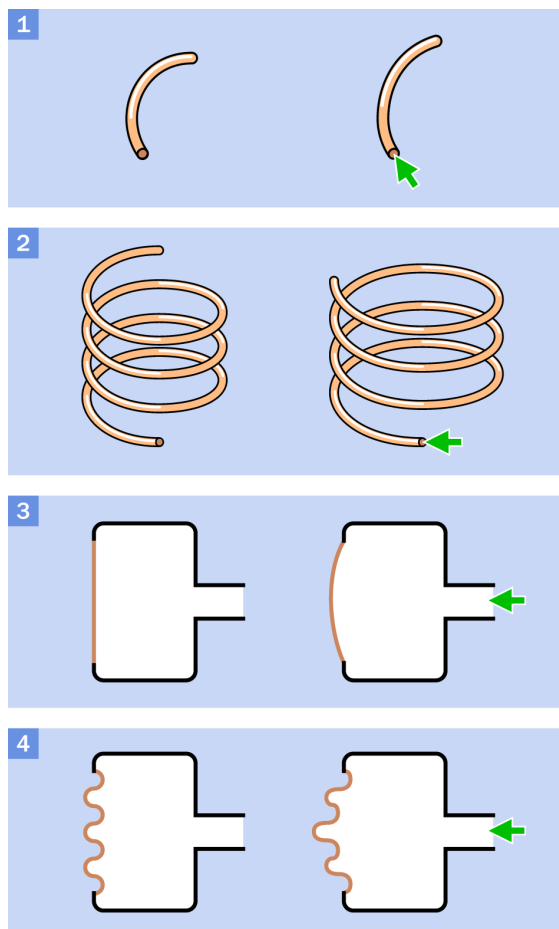


Figure 17-1 Sensing elements. See text for details.

The sensing elements numbered 1 and 2 are less likely to be used in modern systems, although a coiled Bourdon tube may be used to turn a potentiometer in low-cost oil-pressure sensing designs. Number 3 is well suited to MEMS devices, as it can be etched into silicon using the principle illustrated in Figure 17-2.

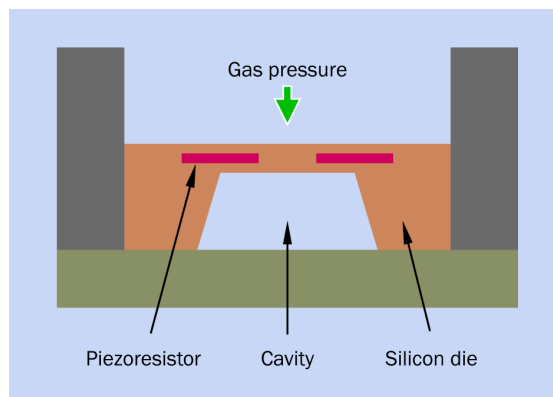


Figure 17-2 A section of a flat-diaphragm pressure sensor, viewed from the side. An aperture in the top face of the chip allows air to enter. The sensor is etched into silicon, and the deflection of the wafer is measured using embedded piezoresistors.

Relative Measurement

Pressure is a relative measurement. It is expressed relative to a *reference pressure* of some kind. Three types of measurement are commonly used:

1. *Absolute pressure*, relative to the zero value of a vacuum.
2. *Gauge pressure*, relative to ambient pressure (that is, pressure in the environment around the sensor). Air pressure is the reference source for gauge pressure, and a *vent* is incorporated into the sensing system.
3. *Differential pressure*. In this case, the pressure being measured is relative to some other pressure—for example, the pressure differential between two sealed tanks.

Figure 17-3 illustrates these three measurement types.

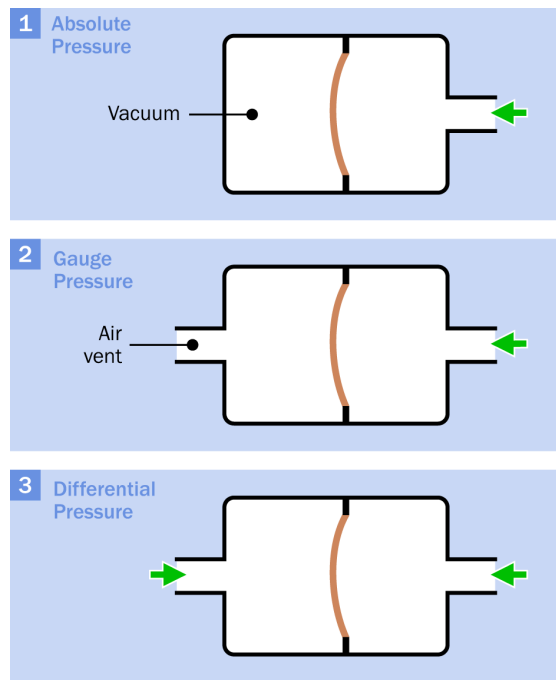


Figure 17-3 Three ways in which pressure is measured.

Variants

Ambient Air Pressure

A *barometric* sensor measures the pressure of air around it. This is an absolute value, relative to a vacuum.

The earliest barometers consisted of a tube sealed at one end, containing mercury. No air was allowed to enter the tube. It was inverted so that its open end was immersed in a small reservoir that was open to the air, the pressure of which supported the column of mercury in the tube. Thus, the height of the column was ratiometrically equivalent to atmospheric pressure. Because low air pressure is often an indicator of inclement weather, a barometer was a very simple tool for weather forecasting.

A modern barometric sensor consists of a chip with a vent hole in its upper surface, allowing ambient air pressure to reach a sensor inside the chip. A popular example has been the Bosch BMP085, shown soldered onto a break-

out board in [Figure 17-4](#). The BMP180 has a very similar specification, the primary difference being that it is designed for a SPI bus instead of I2C. For additional details about protocols such as SPI and I2C, see [Appendix A](#).

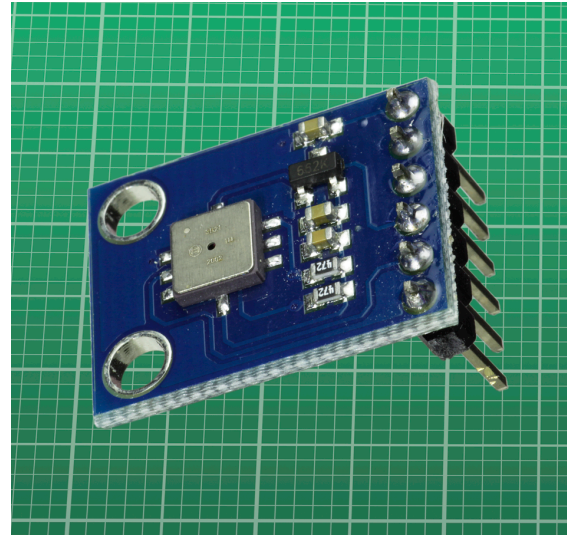


Figure 17-4 A Bosch barometric sensor on a breakout board. The background grid is in millimeters.

This sensor uses a supply voltage ranging from approximately 2VDC to 3.5VDC, but a breakout board includes a voltage regulator that tolerates a 5VDC supply. The output is digitized for access by a microcontroller, but the format is quite complex and consists of raw data that must be converted to air pressure by applying a formula. The manufacturer offers free code in the C language for this purpose. A thermometer that is built into the chip enables temperature compensation.

Subsequent products from Bosch include the BME280, which adds a humidity sensor. Breakout boards using this and other barometric chips are available from sources such as Sparkfun and Adafruit, where Arduino code libraries to interpret the data are available for download.

Altitude

A barometric sensor can be used to determine altitude. At sea level, air pressure is approxi-

mately 101kPa (kilopascals), or 760 millimeters of mercury. At 5,000 meters altitude, air pressure drops to 56kPa. The transition is nonlinear, changing most rapidly at lower altitudes.

Air pressure will be affected by temperature and weather conditions. Weather is unlikely to affect the value by more than plus-or-minus 1%, and temperature has a smaller effect; but the barometric sensor can be reset to a zero point based on current temperature and weather data.

Gas Pressure

Thousands of board-mountable gas pressure sensors are available from retail component suppliers. The need to make connections via tubing imposes a minimum size limit on these parts. Many have a through-hole format, but even the surface-mount variants can be relatively large (e.g., 10mm × 10mm). Most tend to have barbed tubing ports, the “barbs” being ridges to retain push-on flexible tubing. Output may be analog or digital, the digital protocol being designed for I2C, SPI, or plain TTL serial buses.

These components are designed for pressures ranging up to 500kPa (kilopascals). Some manufacturers use psi in their specifications, while others use bars and millibars, and a few state gas pressures equivalent to inches of water.

Dual-ported gas sensors for differential pressure measurement can be used for gauge measurement (i.e., relative to ambient air) if one port is left open. Absolute pressure sensors have only one port, as they measure pressure relative to a vacuum maintained inside the chip, behind the diaphragm.

- Some gas pressure sensors can do dual duty as liquid pressure sensors, but others cannot, and datasheets must be consulted carefully, as online vendors may not include this information prominently.

The sensor in [Figure 17-5](#) by All Sensors Corporation is designed for flexible tubing with 1/16” internal diameter (1.6mm). Its pin spacing is 0.1” (2.54mm), making it breadboardable. This sensor is intended for gases only, but has a “10-inch” water-equivalent rating, meaning that it can tolerate a maximum pressure equivalent to that of a 10-inch (25.4cm) column of water.

When supplied with up to 16VDC, an internal Wheatstone bridge circuit provides an analog output that varies ratiometrically by a few millivolts over the full range of measurable pressures (the exact specification depending on the power supply). The output may be amplified with an external op-amp. Although this is a differential sensor, it can provide gauge pressure if one of the ports is left open.

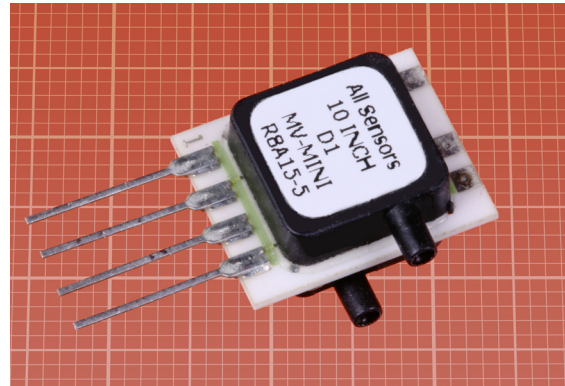


Figure 17-5 A gas pressure sensor for use with 1/16” (1.6mm) internal diameter flexible tubing. The background grid is in millimeters.

The larger sensor in [Figure 17-6](#) is in the ADCA range from All Sensors Corporation, designed for flexible tubing with 1/8” internal diameter (3.2mm). This is intended for gases only, and has a “5-inch” water-equivalent rating. It requires a 5VDC power supply and has an internal op-amp that provides an output that varies ratiometrically by 0.2VDC (centered on 4VDC) over the range of measurable differential pressures. Like its smaller cousin, it can provide gauge pressure if one of the ports is left open.

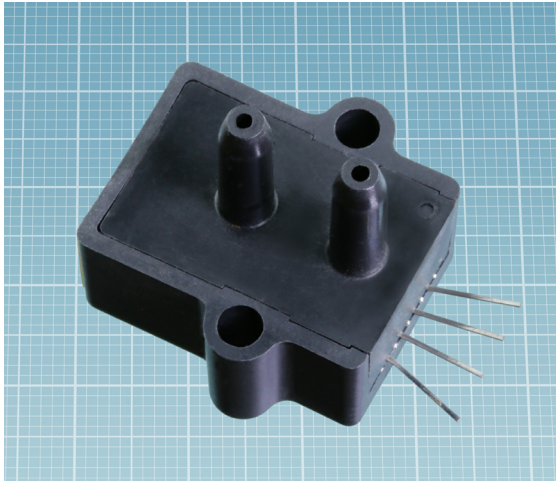


Figure 17-6 A gas pressure sensor for use with 1/8" (3.2mm) internal diameter flexible tubing. The background grid is in millimeters.

What Can Go Wrong

Vulnerability to Dirt, Moisture, and Corrosive Materials

MEMS pressure sensors containing very delicate, very small sensing elements must inevita-

bly come into direct contact with gases and, in some cases, liquids. Datasheets will provide warnings regarding humidity and corrosive fluids, but the risk of contamination with dirt remains. Liquids should be filtered to minimize this risk.

Barometric sensors have a hole in the package, exposing the sensor element to the ambient air. This vent must be protected from direct contact with the environment.

Light Sensitivity

If light is allowed into the vent hole in a barometric sensor, it can create photocurrents in the chip. This will affect accuracy.

gas concentration sensor

Humidity sensors and *vapor sensors* are included in this entry, as they measure concentrations of a liquid while it is in its gas phase.

Sophisticated *industrial sensors* are available for accurate gas sensing, but this entry deals almost entirely with lower-cost solid-state sensors that are often classified as *board-mount* components.

OTHER RELATED COMPONENTS

- **gas/liquid pressure** sensor (see [Chapter 17](#))

What It Does

Small semiconductor-based gas sensors provide a low-cost method for detecting specific gases in ambient air. The sensors vary their electrical resistance or capacitance in response to the concentration of the gas. They may be used in conjunction with an alarm system that will sound if gases such as *propane* or *carbon monoxide* exceed a preset level, or if the concentration of *oxygen* drops below a preset level.

Because vapor consists of a liquid in its gas phase, gas sensors can respond to vapor, as in the case of *alcohol sensors*.

A *humidity sensor* measures the amount of water vapor in the air, which can be important in refrigeration, HVAC (heating, venting, and air conditioning), medical equipment, meteorology, and storage rooms where art objects, antiques, or paper archives must be preserved in an environment that is neither too dry nor too moist, and is held at a constant temperature. Humidity control can be important also in climates where growth of mold is a concern. Humidity sensors are also used in automobile climate control and windshield defogging, and

in the storage of food, fabrics, wood products, and medications.

High humidity coupled with high temperature contributes to the decomposition of many substances, while low humidity can cause desiccation. High humidity can also create damage if it causes materials to expand as a result of taking up moisture.

Many methods have been devised to detect individual gases, but semiconductor sensors are now dominant in applications where accuracy and gas-specificity are not crucial.

Schematic Symbol

No specific schematic symbol exists to represent a semiconductor gas sensor.

Semiconductor Gas Sensors

During the development of transistors in the 1950s, engineers noticed that semiconductor p-n junctions were sensitive to the presence of some gases in the atmosphere. This was regarded as a problem, which was solved by encapsulating transistors to prevent their exposure.

During the 1980s, Japanese law required installation of sensors in every home to detect hazardous concentrations of propane gas. This encouraged the development of cheap, long-lasting components that took advantage of semiconductor sensitivity.

Tin oxide is widely used in a variety of solid-state gas sensors. A sintered layer of the compound is deposited on a ceramic substrate in combination with other compounds such as antimony oxide. The granular layer functions as an n-type semiconductor in which electron transfer increases when certain gases are adsorbed among the grains. When the gas concentration diminishes, oxygen atoms displace the gas molecules, and the sensor returns to its original state. It is unimpaired by being activated, and has a life expectancy of at least 5 years during active (powered) use.

Each sensor includes a tiny resistive heater that is necessary for the chemical reaction to occur. Voltage must be applied to two pins that are connected with the heater. Two other pins connect internally with the sensing element. The resistance between these pins will vary with presence of the target gas; thus, this type of component has a resistive output.

It is a feature of semiconductor gas sensors that they tend to suffer from *cross-sensitivity*. That is, one sensor may respond to more than one gas. Manufacturers control this problem to some extent by adding filtering material around the semiconductor element, or by adjusting the proportions of dopants used in the semiconductor. Datasheets should be consulted carefully to see if a sensor may give false positives as a result of other gases that are likely to be present in the area where it will be used.

Figure 18-1 shows an MQ-5 propane sensor from Hanwei Electronics. The component also responds to methane, hydrogen, alcohol, and carbon monoxide, but with much less sensitivity. The manufacturer recommends calibrating

the sensor by using a resistor ranging from 10K to 47K in series with the output resistance.

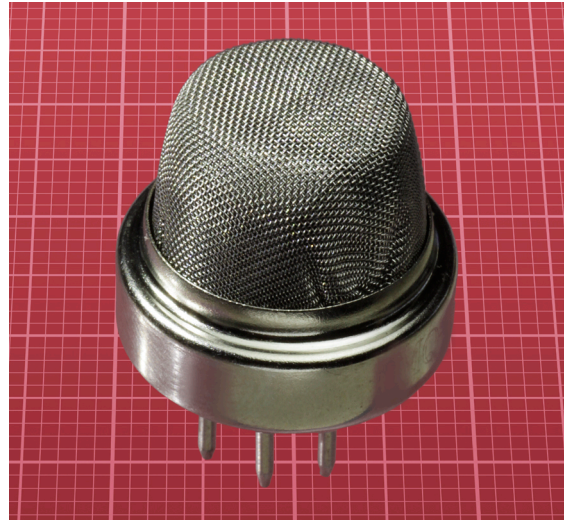


Figure 18-1 A propane gas sensor. The background grid is in millimeters.

Figure 18-2 shows an MQ-3 alcohol sensor, also manufactured by Hanwei. It has some sensitivity to benzene, but this is unlikely to be a problem, as benzene is seldom present in significant concentrations in ambient air. However, the response of the sensor to alcohol varies with temperature and humidity. Consequently this component can only be used as a “breathalyzer” if the user can be satisfied with an approximate response.

Semiconductor gas sensors are also available for detecting methane, carbon monoxide, hydrogen, ozone, and other gases.

These components are not low-current devices. Typically the internal heater in the Hanwei range has a resistance of slightly more than 30 ohms, and will draw 150mA to 160mA at 5V (i.e., slightly less than 1W). Because the heater is a simple resistive device, it can be used with AC or DC. The output resistance of the sensor can also be assessed with an AC or DC signal.

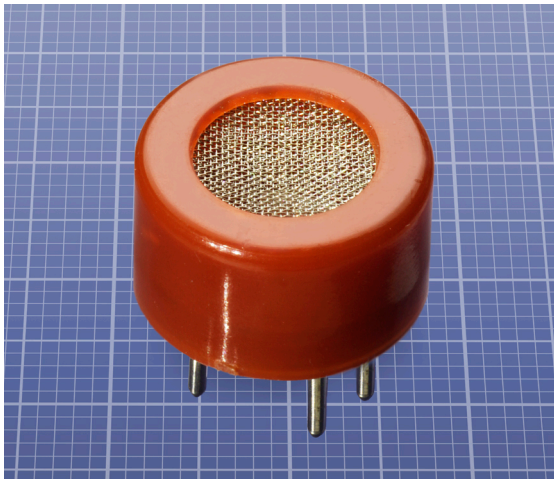


Figure 18-2 An alcohol sensor. The background grid is in millimeters.

A breakout board from Parallax simplifies the use of Hanwei gas sensors. The board is shown in [Figure 18-3](#). It is compatible with carbon monoxide, propane, methane, and alcohol sensors, each of which can be plugged into a socket on the board. Two trimmers establish the sensitivity and the trip point for the sensing element, and the TTL output is then logic-high when a gas is detected, and logic-low otherwise.

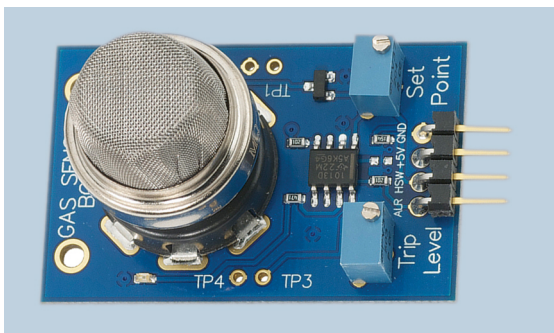


Figure 18-3 A Parallax breakout board to simplify the use of Hanwei gas sensors.

Oxygen Sensors

Oxygen sensors are often built using a membrane made of zirconium dioxide. This material has the property that it can transport oxygen

ions when heated. One setup is to have a zirconium membrane separating the gas to be measured from ambient air. This is a type of fuel cell, called a concentration cell or Nernst cell. If the oxygen concentration differs between the two sides of the membrane, oxygen ions will flow through it. Only oxygen ions can move—not neutral oxygen atoms or molecules. The ions are negatively charged, so the transport will lead to a potential difference over the cell, which can be measured with platinum electrodes.

To comply with emissions regulations, automobiles sense the oxygen level of exhaust gases. This data controls the fuel-air ratio in the fuel injection system. Too much air results in the formation of nitrogen oxides, while too little air results in excessive carbon monoxide.

Humidity Sensors

Moisture content of the air is expressed in three different ways:

Absolute humidity

This is the weight of water vapor in a fixed volume of air. It is measured in grams per cubic meter in the metric system. An absolute humidity sensor is properly called a *hygrometer*.

Dew point

If a sample of air is cooled without a change of pressure, the dew point is the temperature at which moisture will start to condense. The dew point is a way of describing how humid the ambient air is currently, as water will condense more readily in humid conditions.

Relative humidity

This is often referred to by its acronym, RH. If temperature, pressure, and volume of a sample of air remain constant, relative humidity is the ratio between the current value of absolute humidity and the hypothetical level where the addition of

water vapor would result in condensation. The ratio is expressed as a percentage. Thus, if moisture is already condensing in an air sample, relative humidity is 100%. If an air sample contains half the weight of moisture required for condensation to begin, relative humidity is 50%. If there is no moisture at all in the air, relative humidity is 0%.

Colloquial usage of the term “humidity” usually means relative humidity, and sensor output is usually convertible to this value. However, some absolute humidity sensors do exist.

Dew-Point Sensor

Historically, meteorologists used a *chilled mirror hygrometer*, in which a metal mirror was exposed to the atmosphere and cooled until the surface was seen to become misty with condensation. The temperature where this occurred was the dew point.

This system is still used in conjunction with an LED and a **phototransistor**. The LED is positioned so that its light reflects from a mirror, directly onto the phototransistor. The mirror is cooled until mist starts to form, which diffuses the reflected light and causes a well-defined change in output from the phototransistor.

Although the formula linking dew point with relative humidity is complex, a simplified approximation is available that is reasonably accurate so long as the relative humidity is 50% or greater. If RH is the relative humidity, t is the current temperature, and t_D is the dew-point temperature at which mist forms:

$$RH = 100 - (5 * (t - t_D)) \text{ approx.}$$

Although a chilled-mirror dew-point sensor has a reputation for being accurate, it is heavy, expensive, and impractical for most applications outside of meteorology.

Absolute Humidity Sensors

An absolute humidity sensor may use two NTC (negative temperature coefficient) **thermistors** in a Wheatstone bridge circuit. One thermistor is sealed in a compartment containing dry nitrogen, with zero humidity. The other is exposed to the atmosphere. Current passing through the thermistors raises their temperature to at least 200 degrees Celsius. Because heat radiates less efficiently when there is moisture in the air, the exposed thermistor will run hotter, and its resistance will be higher, at higher levels of humidity. This type of sensor is used in clothes dryers and wood kilns, among other applications.

See Chapters 23 and 24 for more information about thermistors.

Relative Humidity Sensors

Two main types of sensing elements are used to measure relative humidity: resistive and capacitive.

In a *resistive* sensing element, a thin layer of polymer, salt, or other hygroscopic substance is deposited on a substrate consisting of ceramic or other unreactive material. When the substance absorbs water, its electrical conductivity increases. Voltage is applied to the sensing element, and AC is used to avoid polarizing it. The current flow is processed externally, with conversion to DC followed by linearization, meaning that the current is processed to establish an almost linear relationship with gas concentration, with temperature compensation factored in. Alternatively, these functions can be performed by hardware built into the sensor, and a digital value for relative humidity can be accessed by an external microcontroller.

In a *capacitive sensor*, once again a thin film of polymer or metal oxide is deposited on a ceramic or glass substrate, but the film functions as a *dielectric* between two metal electrodes that serve as the plates of a capacitor. The dielectric value changes as the film absorbs moisture, and this causes the capacitance to

vary, typically by 0.2pF to 0.5pF for each 1% change in relative humidity. This is almost a linear relationship extending over the entire range from 0% to 100% relative humidity.

The actual capacitance value at 50% relative humidity is likely to be between 100pF and 500pF. The sensor may be excited with AC from an external source, or can be incorporated in a chip that derives AC from a DC power supply and provides a digital output.

To determine the dew point or absolute humidity from a value for relative humidity, ambient temperature must also be measured. A chip such as the Si7005 from Silicon Labs includes a temperature sensor with a relative-humidity sensor based around a capacitor in which polyimide film forms the dielectric. If condensation forms, an on-chip heater will cause it to evaporate so that normal operation can resume. Data from the chip is supplied via an I2C interface.

Humidity Sensor Output

When the output is analog, resistance or capacitance of the internal sensing element is available via two pins or solder pads on the sensor. The analog value must be converted to a value for relative humidity by performing a calculation that takes the current temperature into account. The component may or may not include a temperature sensor.

With a digital output, an internal analog-to-digital converter can be accessed by a microcontroller via serial, I2C, or SPI interface. Alternatively, the sensor may communicate values using pulse-width modulation. Either way, the output is a value for relative humidity, calculated on the chip with reference to an onboard temperature sensor.

Analog Humidity Sensor

The Humirel HS1101 is a low-cost analog-output humidity sensor that varies its internal capacitance between approximately 160pF and 200pF as relative humidity increases from 0% to 100%. The response is almost linear, with the

curve steepening slightly when humidity exceeds 80%. The component is shown in Figure 18-4.

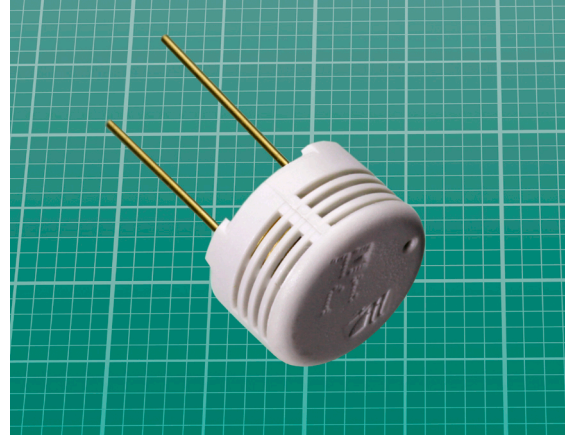


Figure 18-4 A Humirel HS1101 humidity sensor. The background grid is in millimeters.

The manufacturer claims a recovery time of 10 seconds after 150 hours of condensation. In other words, the performance of the sensor should be restored to its original specification.

A microcontroller can evaluate the output by measuring the charge time of the internal capacitor in the sensor. If the sensing element is wired in parallel with a 10M resistor, this will allow the capacitor to discharge before the microcontroller charges it again. A 220-ohm series resistor should be used between the microcontroller and the sensor, to limit the charge current. This circuit is illustrated in Figure 18-5.

Alternatively, the sensor manufacturer suggests using the capacitance value to control the output frequency of a 555 timer. A counter or microcontroller may be used to count the number of pulses per unit of time.

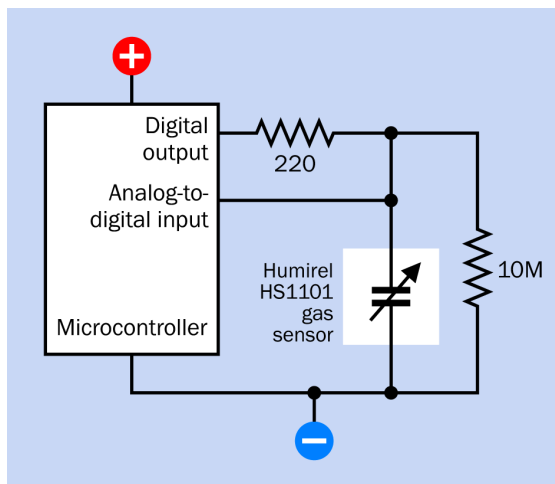


Figure 18-5 Using a microcontroller to measure the charge time of a gas sensor with capacitive output.

The Honeywell HIH4030 is a surface-mount humidity sensor with a more convenient analog voltage output that increases almost linearly from approximately 0.8VDC at 0% humidity to 3.8VDC at 100% humidity, assuming a 5VDC power supply. The sensor is available on a miniature breakout board from Sparkfun, shown in Figure 18-6.

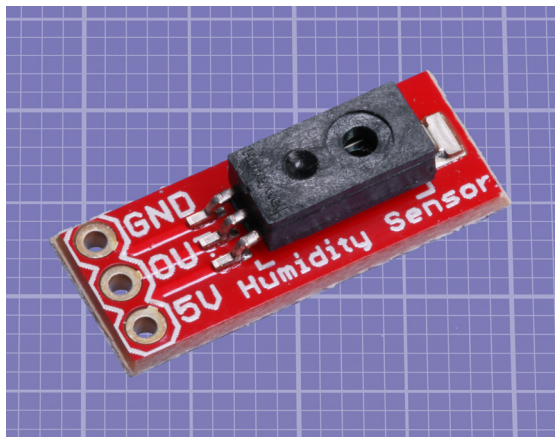


Figure 18-6 The Honeywell HIH4030 on a breakout board from Sparkfun.

Design Considerations

As relative humidity is temperature-dependent, a relative-humidity sensor must have the same temperature as the air it measures. Many data-

sheets recommend that when a sensor is mounted on a printed circuit board, slots should be milled around it to minimize heat transfer to it. The sensor should also be mounted as far away as possible from heat-generating components.

Where a capacitive humidity sensor with an analog output is located some distance from the electronics that will process its output, shielded cables or twisted pair cables should be used to minimize capacitance in the wiring. A decoupling capacitor between the supply voltage and ground close to the sensor can help to keep the supply voltage stable.

Digital Humidity Sensor

The AM2302, available from Adafruit, is a capacitive humidity sensor with a digital output accessible by a microcontroller via the I2C protocol. Onboard electronics calculate relative humidity with reference to an included temperature sensor. This component is pictured in Figure 18-7.

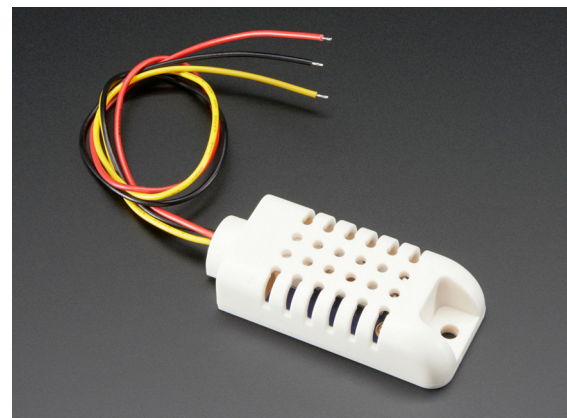


Figure 18-7 A low-cost humidity sensor with temperature-compensated digital output. From a photograph by Adafruit.

What Can Go Wrong

Contamination

A semiconductor gas sensor can be damaged by exposure to volatile chemical vapors. This

would be unusual in a domestic environment, but is still an important consideration, as there will be no obvious indication that damage has occurred.

Recalibration

If a humidity sensor is exposed to very high humidity where condensation occurs, some

datasheets advise a baking procedure, where the sensor is placed in warm, dry air for several hours and then allowed to rehumidify.

Soldering

Semiconductor gas sensors should be soldered quickly and at a controlled temperature, to minimize heat transfer.

gas flow rate sensor

19

Sensors that contain no electronic components are not included in this entry.

A **gas flow rate** sensor may often be described as a *mass flow* sensor or *mass flow rate* sensor. Although it measures volume rather than mass, the mass of gas can be calculated so long as its temperature and pressure are controlled.

Some methods of gas flow sensing can also be applied to liquids, but sensors are usually designed for one application or the other. Therefore, **liquid flow rate** sensors have their own entry. See Chapter 16.

An *anemometer* is a gas flow rate sensor that measures air speed. It is included in this entry.

Many gas flow rate sensors are large devices designed for industrial applications. This entry focuses mostly on lower-cost solid-state sensors that are often classified as *board-mount* components.

OTHER RELATED COMPONENTS

- **liquid flow rate** sensor (see Chapter 16)

What It Does

A **gas flow rate** sensor measures the volume of gas flowing past or through the device, usually inside a pipe. In most applications, users wish to know the mass of the gas that is passing per unit of time. Consequently a gas flow rate sensor is very often identified as a *mass flow rate* sensor. If it functions by heating the gas and measuring the heat dissipation, it is a *thermal mass flow rate sensor*.

A sensor that measures the flow of open air is often described as an *anemometer*. Its output will be expressed as a velocity, not as a volume or mass.

Applications

Mass flow rate sensors are used frequently for laboratory and medical applications, although the reliability and affordability of thermal mass flow rate sensors is now making them attractive for metering municipal gas supplies.

Anemometers are used mostly in meteorology, aviation, and on boats.

Schematic Symbol

Many specialized symbols are used in flow diagrams to represent pumps, valves, and sensors, but they are not schematic symbols of the type generally found in electronic circuit diagrams, and are not included here.

How It Works

Because an anemometer functions so differently from a gas flow rate sensor that is enclosed in a pipe, the two types of sensors will be discussed separately.

Anemometer

The name of this device is derived from the Greek word *anemos*, meaning “wind.” The *cup anemometer* was invented in 1846, using four hemispherical cups attached to horizontal arms rotating on a vertical shaft. The rate of rotation was proportional to wind speed over a substantial range, but the conversion factor between wind speed and RPM varied depending on the size of the cups and their distance from the shaft.

Anemometer design was simplified to three cups in 1926 and was modified in 1991 by adding a tag to one cup. This causes the speed of rotation to fluctuate as the tag rotates through the wind stream, and the direction of the wind can be calculated from the speed fluctuations. Not all anemometers rely on this principle, however, and a separate wind vane can be used to determine wind direction.

The basic design of a modern anemometer is illustrated in Figure 19-1.

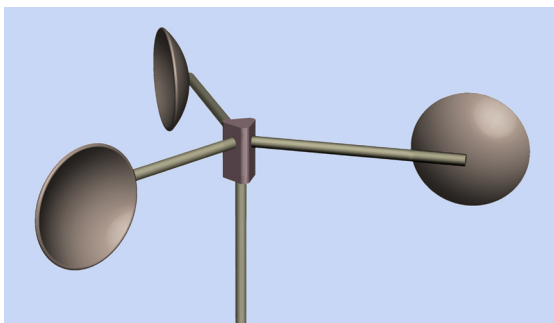


Figure 19-1 The basic design of a meteorological cup anemometer.

Anemometers traditionally used a mechanical counter to log the number of rotations, which were checked at fixed intervals to derive the

wind speed. The output of a modern anemometer may be achieved by generating AC or DC power, or from *Hall-effect sensors* (see “*Hall-Effect Sensor*” for a full discussion of Hall-effect sensors).

Handheld Anemometer

A digital handheld anemometer for personal use is shown in Figure 19-2. A cup anemometer made by Vaavud for use with a mobile phone (with appropriate software) is shown in Figure 19-3.



Figure 19-2 A handheld digital anemometer.

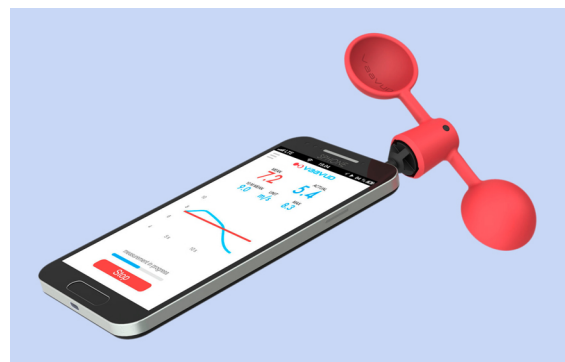


Figure 19-3 A cup anemometer sold as an accessory for smartphones.

Ultrasound Anemometer

The movement of air affects the speed of sound, enabling calculation of both wind direction and wind velocity by using an array of

ultrasound emitters and receivers. The lack of rotating parts promises greater reliability, and in Figure 19-4 an ultrasound anemometer manufactured by Biral Metereological Sensors also includes heaters so that it will be immune to snow or ice accumulation during freezing conditions.

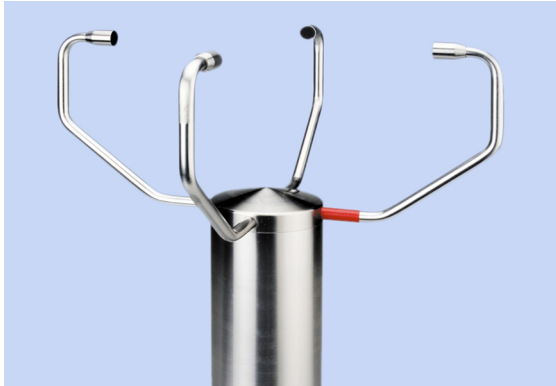


Figure 19-4 Made by Biral Metereological Sensors, this anemometer uses ultrasound to determine wind speed and direction.

Ultrasound anemometers have been built on a DIY basis by hobbyists, typically using off-the-shelf ultrasound emitters and an Arduino to decode the signals. Several of these projects are documented online.

Hot Wire Anemometer

A hot wire anemometer measures the cooling effect of the air. It heats a thin wire by passing current through it, and measures the heating power needed to keep the temperature constant.

It is also possible to keep either the voltage over the wire or the current constant, and to assess the wire temperature. The temperature can be measured directly or calculated from the wire's resistance, which increases as the wire gets hotter.

Mass Flow Rate Sensing

A mass flow rate sensor measures the flow speed of a gas. When this is multiplied by the density, the mass flow rate can be calculated.

Most sensors of this type heat the gas and are categorized as *thermal mass flow rate* sensors. The gas passes over a *thermopile* (consisting of several thermocouples wired in series), then a heater, and then another thermopile. These components are miniaturized and can be etched into a chip measuring 2mm square or less.

The temperature difference between the two thermopiles increases as the flow of gas becomes faster and transports more heat to the second thermopile. This is known as the *heat transfer principle*, illustrated in Figure 19-5.

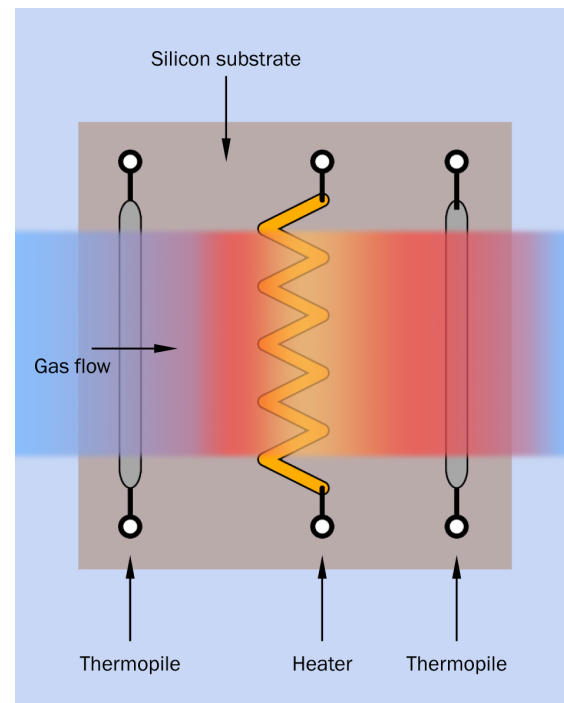


Figure 19-5 In a thermal mass flow sensor, the flow of gas heats the thermopiles disproportionately, and the temperature difference can be used to evaluate the rate of flow.

This principle is used also in a liquid thermal mass flow rate sensor, described in “*Thermal Mass Liquid Flow Rate Sensor*”.

An example of a thermal mass flow rate sensor is shown in Figure 19-6.



Figure 19-6 A thermal mass flow rate sensor made by the Chinese manufacturer Zhengzhou Winsen Electronic Technology Co. Ltd.

Applications

In medicine, mass flow rate sensors are used in anesthesia delivery, respiratory monitoring, sleep apnea machines, ventilators, and other devices. Industrial uses include air-to-fuel ratio measurement, gas leak detection, and gas metering.

While the metering of municipal gas supplies was traditionally done with all-mechanical devices, MEMS-based meters are replacing many of the 400 million mechanical gas meters estimated to exist worldwide.

Units

Mass flow rate sensors are often rated in SLM, meaning standard liters per minute. A “standard liter” has a temperature of 0 degrees Celsius and a pressure of 101.325 kPa (kilopascals). This pressure is equivalent to that of air at sea level. Because temperature and pressure are specified, the mass of a gas in a standard liter can be calculated by knowing the density of the gas. Thus SLM is a way of measuring mass flow, even though it is a measurement of volume.

The acronym SLPM is also used, but it means the same thing as SLM. SLs and SLPs are measurements of standard liters per second, while SCCM refers to standard cubic centimeters per minute.

Measuring Higher Volumes

MEMS sensors are typically equipped with inflow and outflow nozzles suitable for flexible tubing of 3mm or 5mm internal diameter. The nozzles are “barbed,” meaning that they have ridges to retain the tubing.

Small tubing can only deal with low flow rates. A few sensors are threaded for standard plumbing pipes, and can accept volumes of up to 10 liters per minute. They are a minority. A low-rate sensor can still be used to measure higher volumes if it is supplied with just a percentage of the primary flow. This principle is illustrated in [Figure 19-7](#), where an adjustable vane in the main pipe creates a pressure differential. A narrower, constricting section of pipe could have a similar effect, but it would not be adjustable.

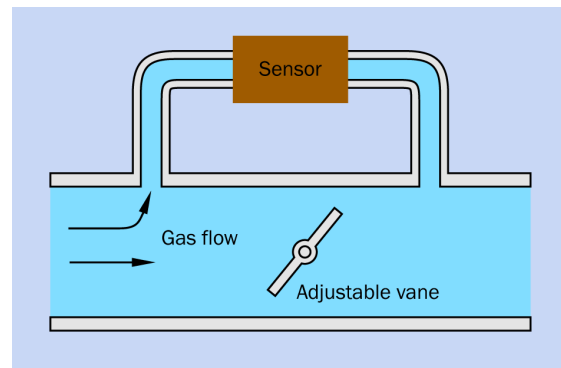


Figure 19-7 A vane in a primary pipe can divert a percentage of the flow to a sensor.

Output

Many mass flow sensors have an analog output consisting of a voltage that varies ratiometrically with gas flow. With a typical 5VDC power supply, output voltage may vary from around 1VDC to 4VDC.

Some sensors now incorporate analog-to-digital converters and data processing to provide SLM digital values, accessible from a microcontroller via an I2C interface.

What Can Go Wrong

The primary risk for gas flow rate sensors is damage caused by particles and contaminants in the gas stream. A 5-micron filter is recommended. A *dust segregation system* can also be

used, consisting of a small compartment containing semicircular centrifugal chambers. Dust tends to follow the outer edge of the flow path, while the flow sensor is placed on the inner side of the path.

photoresistor

20

A **photoresistor** has a function similar to that of a **phototransistor**, but as its name implies, it is a purely passive component that varies its resistance in response to light.

The term *photocell* was used formerly, but has been displaced by the term **photoresistor**, which describes its function more accurately. The term *photoconductive cell* is sometimes used, or *light-dependent resistor* (or its acronym, *LDR*).

OTHER RELATED COMPONENTS

- **resistor** (see Volume 1)
- **photodiode** (see [Chapter 21](#))
- **phototransistor** (see [Chapter 22](#))

What It Does

Formerly known as a *photocell*, a **photoresistor** is a disc-shaped component with two leads. When light falls on the surface of the disc, resistance between the leads will diminish. Some photoresistors have a resistance in darkness as high as 10 megohms. A few have a resistance in bright light as low as 500 ohms, although several kilohms would be more common.

A photoresistor is less sensitive to light than a **phototransistor** or **photodiode**, and unlike them it is a passive component with no polarity. It presents equal resistance to current in either direction, and may be used with DC or AC.

Because cadmium sulfide is commonly used in this component and is regarded as hazardous to the environment, photoresistors are now unavailable in some regions (notably, Europe). However, at the time of writing, they are still

available from many Asian sources, and from some importers in the United States.

Schematic Symbol

Six schematic symbols for a photoresistor are shown in [Figure 20-1](#). They are functionally identical, regardless of whether the single slanting arrow across the resistor symbol is omitted.

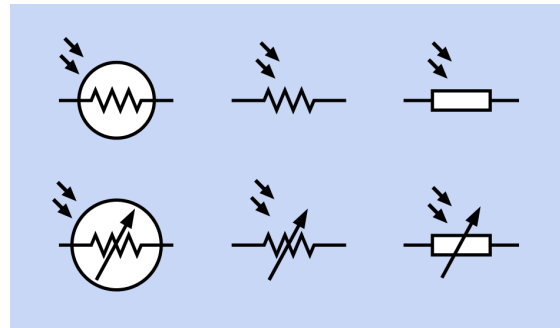


Figure 20-1 All six symbols for a photoresistor are functionally identical.

How It Works

Cadmium sulfide is a semiconductor. When exposed to light, more charge carriers are excited into states where they are mobile and can participate in conduction. As a result, electrical resistance decreases.

Construction

A closeup of a photoresistor appears in [Figure 20-2](#). The brown material is a layer of cadmium sulfide deposited onto a ceramic base. The silver material is a conductive compound evaporated onto the cadmium sulfide to form two interlocking electrodes, in a pattern that maximizes the length of the boundary between each of them and the semiconductor. The electrodes connect with leads projecting from the back of the component.



Figure 20-2 Closeup of a photoresistor. Two interlocking electrodes are mounted on a brown semiconductor layer.

Variants

[Figure 20-3](#) shows a variety of photoresistors, illustrating the range of sizes available. Small photoresistors may be less than 5mm in diameter; large ones may be 25mm in diameter. The size generally suggests the ability of the component to pass current.

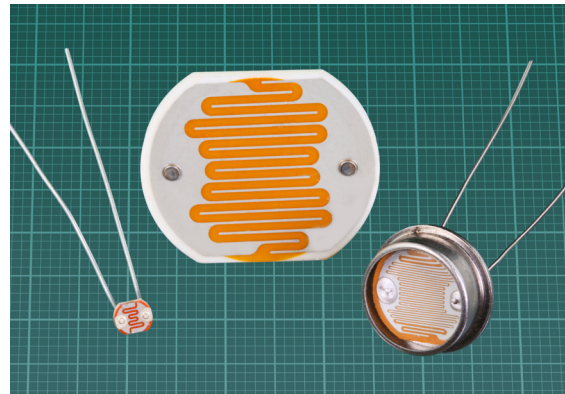


Figure 20-3 Photoresistors are available in a wide range of sizes. The component in the center is shown at the same scale as the others. Generally speaking, larger components are able to conduct higher currents. The background grid is in millimeters.

Photoresistors in Optical Isolators

An *optical isolator*, popularly known as an *opto-coupler*, contains an LED opposite to a photoresistor, in a sealed package. This is discussed in Volume 2. A *Vactrol* is a similar component, except that the LED is placed opposite to a photoresistor. An example is shown in [Figure 20-4](#).

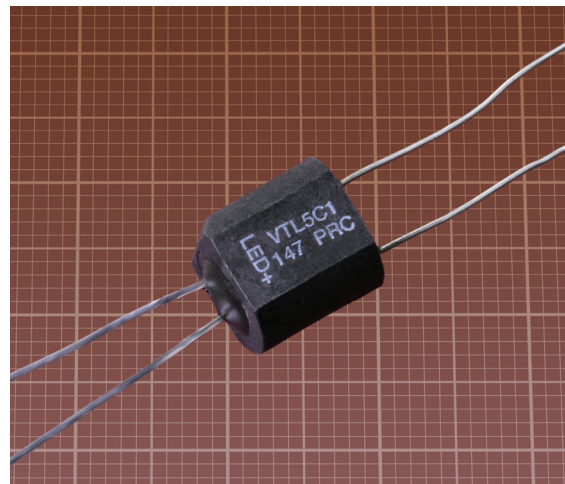


Figure 20-4 A Vactrol, containing an LED opposite to a photoresistor. The background grid is in millimeters.

Vactrol is a brand name owned by its initial manufacturer, Vactec. It was developed to con-

trol a vacuum tube, hence its name. In the 1950s, Vactrols were used in guitar amplifiers to control tremolo.

Because of the relatively slow response time of a photoresistor, and its sensitivity to temperature, optical isolators based on photoresistors are not used in digital devices. They still retain some utility in audio components and music equipment, where the ability of the photoresistor to pass AC is an advantage and its response time is adequate.

Values

Datasheets for a few photoresistors can still be found from some suppliers, such as Digi-Key, but are mostly unobtainable, as major semiconductor companies have stopped making these components. Vendors may quote basic values, but in the absence of part numbers or a manufacturer name, a buyer cannot verify the information.

The resistance range can be determined by testing a sample component. A typical range would be from 50K in a light intensity of 10 lux, up to 1M in darkness. Maximum power dissipation is likely to range from 100mW for a small photoresistor to 500mW for a large one.

Maximum voltage may be as high as 200V, but photoresistors will work just as well at low voltages.

Comparisons with a Phototransistor

Slower response

A photoresistor typically takes several milliseconds to respond to bright light, and can require more than 1 second to regain its dark resistance. A phototransistor is much more responsive, and a photodiode is faster still.

Narrower range of resistance

The maximum resistance of a photoresistor is almost always significantly lower

than the effective maximum resistance of a phototransistor in darkness, and the minimum resistance is likely to be significantly higher than that of a phototransistor in bright light.

Greater current-carrying capacity

Often a photoresistor is rated for twice as much current as the output from a phototransistor.

Not directional

Because a photoresistor is not packaged with a lens, it is sensitive to incident light from anywhere in front of it. If an application requires that light sensitivity should be confined within a narrow angle while the component is insensitive to incident light from other directions, a phototransistor or photodiode should be used.

Temperature dependent

The resistance of a photoresistor varies more with temperature than the effective resistance of a phototransistor.

Cost

At this time, photoresistors are likely to be more expensive than phototransistors.

Lack of information

Photoresistors are often sold without means of checking their specification in a datasheet.

How to Use It

While the effective resistance of phototransistors and photodiodes varies with the applied voltage, photoresistors present the same resistance for a particular light intensity regardless of the voltage applied. This property has made them suitable for use in “stomp box” guitar-effects pedals.

Because the minimum resistance of a photoreistor in bright light tends to be relatively high, and because its response time is quite slow, it is

suitable primarily as an analog component rather than as a switch.

In principle, any resistor in a circuit can be replaced by a photoresistor, and its smooth response to light variations would make it suitable, for example, as the resistance in an RC oscillator circuit, where it would determine the charge time of a capacitor. The frequency of the circuit would thus become light-dependent.

Cadmium sulfide photoresistors respond most actively to wavelengths of light ranging from 400nm to 800nm. This is especially important where an **LED indicator** is used as a light source, as LEDs often emit an extremely narrow range of wavelengths. (See the entry for LED indicators in Volume 2.)

Choosing a Series Resistor

To convert light intensity into a voltage, a photoresistor can be connected in series with a regular resistor, to form a voltage divider. There are two ways to do this, as shown in Figure 20-5. On the left side of this figure, light falling on the photoresistor will cause the output voltage to rise. On the right side, light will cause the output to drop.

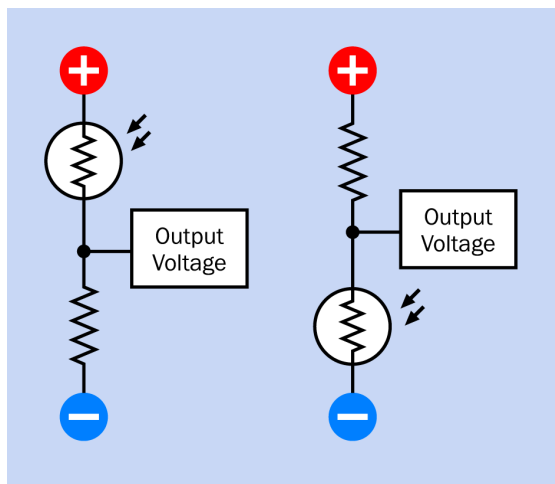


Figure 20-5 Using a photoresistor to create a variable voltage. See text for details.

If R_{MIN} is the minimum photoresistor value in the brightest light that will be used, and R_{MAX} is the maximum value in the dimmest light that will be used, a simple formula can be used to find R_s , the optimum value for a series resistor, which will provide the widest range of voltage values at the center of the voltage divider:

$$R_s = \sqrt{R_{\text{min}} * R_{\text{max}}}$$

What Can Go Wrong

Overload

Because a datasheet may be unavailable, a photoresistor may have to be used on a trial-and-error basis. A test-to-destruction approach may be necessary to determine the limits of the component.

Excessive Voltage

Exceeding the maximum rated voltage for a photoresistor, even for a short time, can cause irreversible damage. Depending on the component, overvoltage can range from 100V to 300V.

Confusion Among Components

Because photoresistors often have no part numbers printed on them and may be sold in miscellaneous assortments, one component may be easily confused with another that has different characteristics. Where two or more photoresistors are used in one device, their characteristics should be measured to determine whether they are functionally identical.

photodiode

21

OTHER RELATED COMPONENTS

- **diode** (see Volume 1)
- **photoresistor** (see Chapter 20)
- **phototransistor** (see Chapter 22)

What It Does

Light falling on a **photodiode** causes it to generate a very small current. This is often called the *photovoltaic effect*. The component functions like a solar panel; in fact, a solar panel can be thought of as being an array of very large photodiodes.

Often, a DC power source is used to apply *reverse bias* to a photodiode. This enables the component to deliver more current. It is now operating in *photoconductive* mode.

Schematic Symbols

Schematic symbols for a photodiode are shown in Figure 21-1. Wavy arrows are often (but not always) used to indicate infrared light.

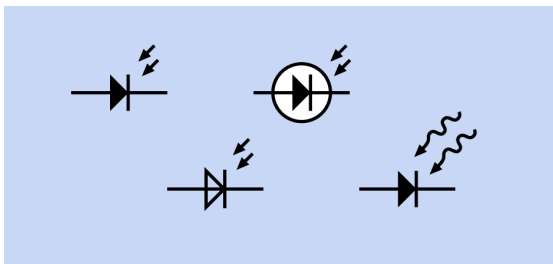


Figure 21-1 Symbols representing a photodiode.

Applications

The rapid response of photodiodes makes them suitable for use in optical disc drives, telecommunications, infrared data transfer, digital cameras, and optical switches. Many sensors in this Encyclopedia use a photodiode. **Proximity** sensors, optical encoders, and light meters are examples.

How It Works

When light falls on a semiconductor, it can excite an electron to a higher energy state. The electron then becomes mobile, and leaves behind an *electron hole* (See the entry discussing **diodes** in Volume 1.)

In photovoltaic mode, incident light creates pairs of electrons and electron holes in the semiconductor material. The electrons move to the cathode of the diode while the holes move to the anode, creating voltage between the two. Note that this happens to some extent even in the absence of visible light, as the photodiode may respond to infrared radiation. The tiny amount of current created without visible light is called *dark current*.

In photoconductive mode, light falling on the photodiode creates pairs of electrons and elec-

tron holes in the semiconductor material. These will move in opposite directions due to the bias voltage, which means a small current flows through the diode.

Photoconductive mode enables a faster response than photovoltaic mode, because reverse-bias voltage makes the depletion layer wider, which in turn decreases the capacitance. (The same effect is used in [capacitance diodes](#).)

Simplified circuits showing the component in photovoltaic mode and photoconductive mode appear in [Figure 21-2](#).

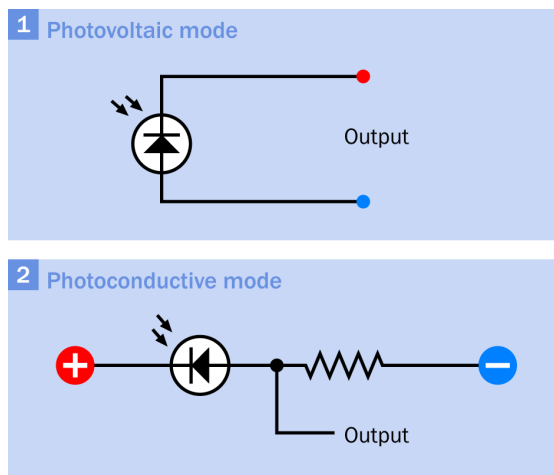


Figure 21-2 Two modes in which a photodiode may be used. In photovoltaic mode, voltage between the two pins must be measured.

Variants

PIN Photodiodes

Like [PIN diodes](#), PIN photodiodes incorporate an undoped (intrinsic) semiconductor layer between the p- and n-doped layers. They are more sensitive and have faster responses than regular PN photodiodes. Many of the photodiodes available are of the PIN type.

Avalanche Diodes

When light enters the undoped region of the avalanche photodiode, it triggers the creation of electron-hole pairs. When electrons migrate

toward the [avalanche region](#) of the diode, the cumulative field strength increases their velocity to the point where collisions with the crystal lattice create further electron-hole pairs.

This behavior causes the avalanche diode to be more sensitive than a PIN photodiode. However, the sensitivity also makes it vulnerable to electrical noise, and it is significantly affected by heat. A guard ring is added around the p-n junction, and a heat sink is often used.

Packages

Many surface-mount and through-hole versions are available. A selection is shown in [Figure 21-3](#) and [Figure 21-4](#). While a through-hole photodiode may look indistinguishable from a through-hole 3mm or 5mm LED, versions are available without lenses, and some are sensitive to incident light coming from the side (they are referred to as [side-looking](#) or [side-view](#) variants).

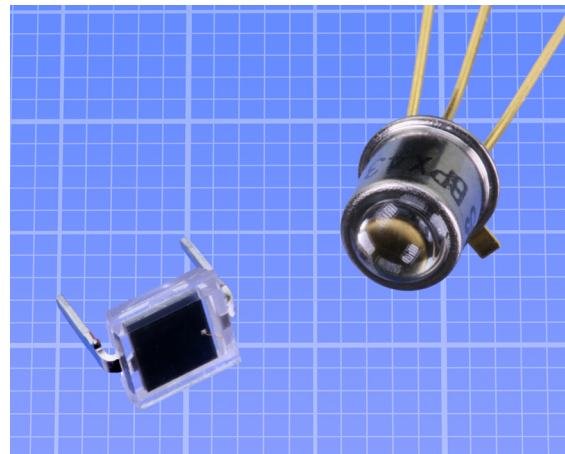


Figure 21-3 Two sample photodiodes. Left: top-view unfiltered Vishay BPW34. Right: Osram BPX43 in metal can suitable for temperatures up to 125 degrees Celsius.

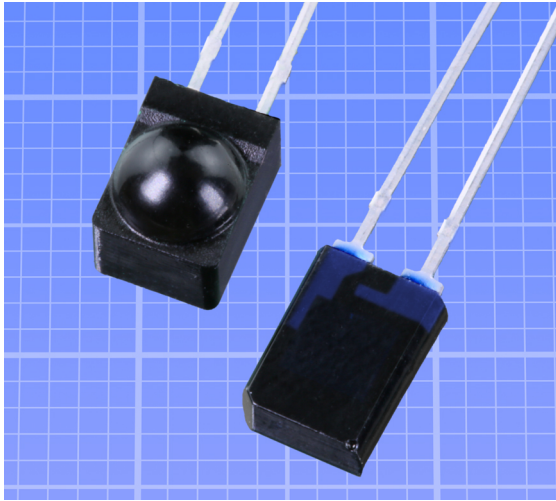


Figure 21-4 Two side-looking photodiodes. Left: Vishay BPV22NF with lens. Right: Vishay BPW83 without a lens. Both have daylight-blocking filters.

Wavelength Range

In order for a photon to be detected by the photodiode, it must carry enough energy to be able to create an electron-hole pair. This energy is a property specific to the semiconductor material, and is known as its *band gap*. Additionally, the epoxy packaging of the photodiode may be designed to block some wavelengths of light. Often, an application will require that the component should only respond to infrared light, not visible light.

Photodiode Arrays

A *photodiode array* has several photodiodes mounted in a row or in a grid, for imaging or measurement applications. A row of photodiodes may be used in a flatbed scanner, where it is moved relative to the reflective object being scanned.

In some arrays, photodiodes are available with color filters preinstalled, to facilitate full-color scanning using the transmitted primary colors.

Output Options

Because the output from a photodiode must be processed to be usable, options exist to convert it to a more convenient form, such as a wider

range of voltages, a square wave signal where the frequency is proportional with light input, or a binary value accessible by a microcontroller via a serial bus such as I2C. For a discussion of sensor outputs see [Appendix A](#).

Specific Variants

Light to Frequency

The Taos TSL235R is a 3-pin, through-hole chip. It combines a photodiode with logic that creates a square-wave pulse train in which frequency is proportional to the light intensity.

Logarithmic Light Meter

The Sharp GA1A1S202WP light sensor has an output voltage that changes logarithmically with the light level. This gives the sensor a large dynamic range, from 3 lux to 55,000 lux, without requiring a high-resolution analog-to-digital converter. (Human perception of light and sound levels is approximately logarithmic.) This is a surface-mount chip, but is available on a breakout board from Adafruit.

Ultraviolet to Analog

The ML8511 from Lapis Semiconductor combines an ultraviolet-sensitive photodiode with an op-amp that provides an output voltage from approximately 1V to 3V, varying with ultraviolet intensity. A breakout board containing this surface-mount chip is available from Sparkfun and many other sources.

Ultraviolet to Digital

The SI1145 from SiLabs combines ultraviolet sensing with data processing to create a UV index, readable from a microcontroller with I2C. Adafruit offers it on a breakout board.

Color to Digital

The Taos TCS3414FN module contains photodiodes sensitive to red, green, blue, and clear (no filtering). Four 16-bit analog-to-digital converters, one for each

channel, provide a digital output accessible over an I2C bus. This module can be used to determine the color of ambient light with some accuracy.

Color to Analog

The Taos TCS3200 also uses red, green, blue, and clear photodiodes but encodes the output from each as a square wave in which the frequency corresponds with the light intensity. The surface-mount chip is available on a breakout board from Robot Shop, shown in [Figure 21-5](#).

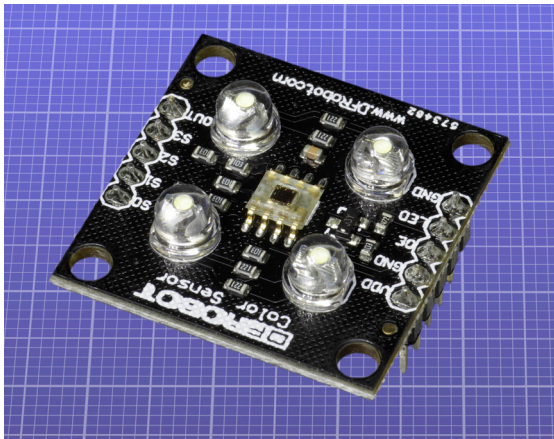


Figure 21-5 This breakout board from Robot Shop uses the Taos TCS3200 chip to analyze the color of incident light.

Values

Abbreviations found in datasheets are included in the list below, with values in parentheses for an Osram SFH229FA infrared photodiode, which resembles a 3mm through-hole LED. It has a peak sensitivity of 880nm and appears black to the human eye, being opaque to wavelengths of light shorter than 700nm, the red end of the visible spectrum.

In [Figure 21-6](#) the SFH229FA is shown beside the SFH229, which has the same peak sensitivity of 880nm but is encapsulated in an untinted module, allowing a sensitivity that tapers gradually to below 400nm, in the green part of the

visible spectrum. With the exception of their spectral range, the two photodiodes have identical specifications.

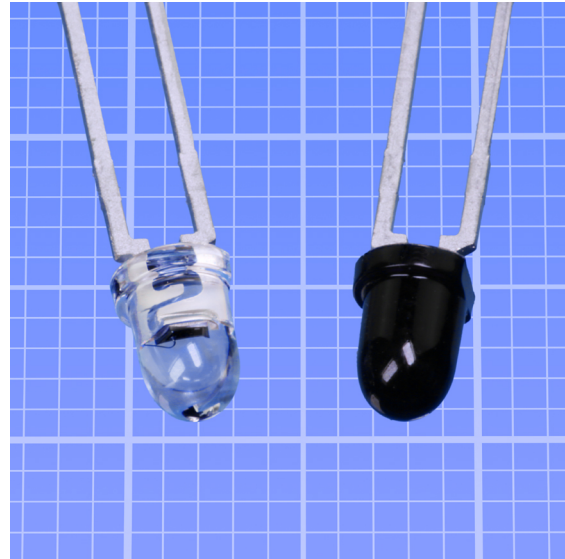


Figure 21-6 Osram infrared photodiodes SFH229 (left) and SFH229FA (right). The background grid is in millimeters.

- Typical forward voltage: V_F (1.3V)
- Typical photocurrent: I_P (20 μ A)
- Maximum power dissipation: P_{TOT} (150mW)
- The *half angle* is measured from the axis of the photodiode to the angle at which the sensitivity has dropped by 50% (plus or minus 17 degrees)
- Dark current: I_R (50pA)
- Wavelength of maximum sensitivity: $\lambda_{S_{MAX}}$ (880nm)
- Response speed is the rise and fall time of photocurrent: t_r and t_f (10ns)

Infrared photodiodes exist with a variety of peak wavelengths. They are designed to function in conjunction with an LED that has a matching wavelength.

The angle of sensitivity depends on the geometry of the package.

Rise and fall speeds are important for high-speed measurement, signaling, or data transfer. The rise and fall time of a typical photodiode can be 1,000 times faster than that of a phototransistor. See “Values” in Chapter 22 for a comparison. Note also that the dark current of a photodiode is much lower than that for a phototransistor.

How to Use It

In photoconductive mode, the photodiode can be connected in series with a suitable resistor so that a voltage divider is formed, as shown in Figure 21-2. The voltage at the output will then vary almost linearly with the intensity of the light.

In the photoconductive mode of operation, the output signal is generally measured in millivolts and microamperes. This signal needs to be amplified, usually with an **op-amp** (described in Volume 2).

Figure 21-7 shows simplified schematics for a standard voltage amplifier, in section 1, and a transimpedance amplifier, in section 2.

A transimpedance amplifier measures the current through the photodiode and converts it into a voltage, without the need for a voltage divider. Advantages include less noise, and no need to determine the value of a voltage-divider resistor.

The output voltage of this amplifier is calculated by using this simple formula:

$$V = R * I_p$$

R is the value of the feedback resistor, which determines the gain of the amplifier.

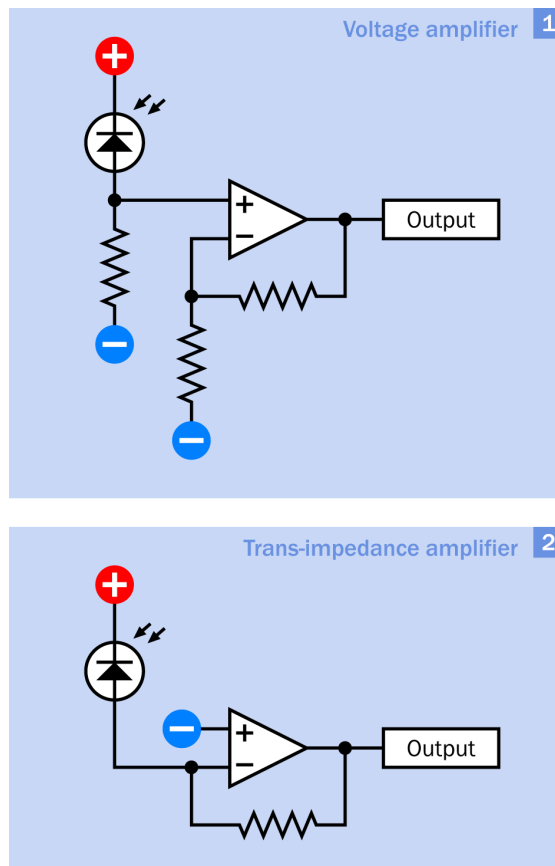


Figure 21-7 Simplified op-amp circuits for use with photodiodes.

What Can Go Wrong

Photodiodes can be hard to distinguish from phototransistors and LEDs, in particular infrared ones. They are typically not marked with type numbers. Using a meter may not be helpful, as regular LEDs behave similarly to photodiodes.

Decision procedure:

1. Does the component conduct in either direction while shielded from light? If so, it is a diode and not a phototransistor or photodiode.
2. Pass a weak current (e.g., 4mA) in the forward direction. If it emits visible or infrared light, it's an LED. (Infrared light may be visible when using a digital camera, or can be detected with a known phototransistor or photodiode.) If the package is clear but cloudy so that the die cannot be seen, it is probably a white LED in which the cloudiness is the fluorescent pigment converting blue light into white.

phototransistor

22

OTHER RELATED COMPONENTS

- **photodiode** (see [Chapter 21](#))
- **photoresistor** (see [Chapter 20](#))
- **passive infrared** sensor (see [Chapter 4](#))
- **transistor** (see Volume 1)

What It Does

A **phototransistor** is a transistor controlled by exposure to light. (**Transistors** are described in Volume 1.) It can be either a bipolar transistor or field-effect transistor (FET), and its body is often superficially similar in appearance to a 3mm or 5mm **photodiode** or **LED indicator** encased in resin or plastic. (LED indicators are described in Volume 2.) However, some phototransistors are encased in a metal shell with a window in it.

The window or the plastic body is either transparent to visible light, or may appear black if the component is intended for use only with infrared while blocking visible wavelengths. A selection of phototransistors is shown in [Figure 22-1](#) (left: Optek/TT Electronics OP506A with a broad spectral response centered around 850nm; center: Vishay TEKT5400S with a side-view lens; right: Vishay BPW17N).

Typically a phototransistor has two leads that connect internally with its collector and emitter (or source and drain, in the case of an FET). The base of the transistor (or the gate of an FET) responds to light and controls the flow of current between the leads.

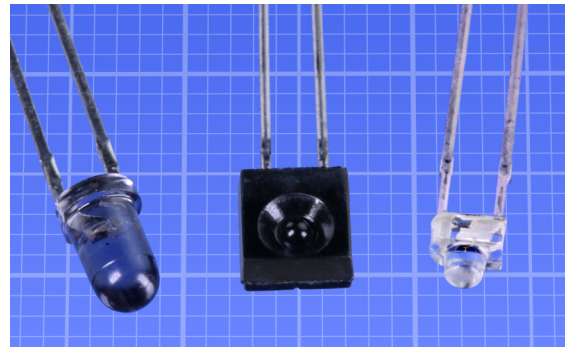


Figure 22-1 A variety of phototransistors. The background grid is in millimeters. See text for details.

In the absence of light, a bipolar phototransistor permits leakage between collector and emitter of 100nA or less. When exposed to light, it conducts up to 50mA. This alone differentiates it from a photodiode, which cannot pass much current.

Schematic Symbols

Symbols for a phototransistor are shown in [Figure 22-2](#). They are functionally identical, with the exception of type C, where an additional connection to the base is included. Often (but not always), wavy arrows indicate infrared light.

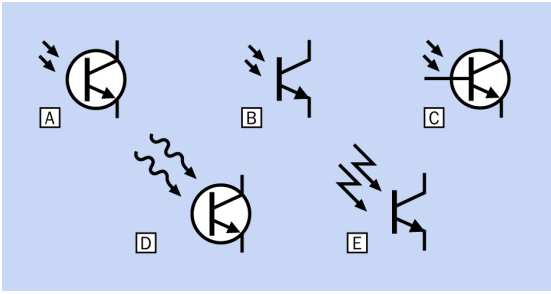


Figure 22-2 Schematic symbols for a phototransistor.

Applications

A phototransistor may be used for light measurement or as a light-sensitive switch.

Often the output from a phototransistor is used in conjunction with a [microcontroller](#) containing an analog-to-digital converter.

Where a clean on-or-off signal is required, a phototransistor can drive the input of a logic chip that contains a [Schmitt trigger](#), or it can be processed by a **comparator**.

An [optocoupler](#) or **solid-state relay** (described in Volume 1) usually contains a phototransistor that is activated by an internal LED. Its purpose is to switch current while electrically isolating one section of a circuit from another.

How It Works

Like a photodiode, the phototransistor responds to light when the light generates electron-hole pairs in the semiconductor material. For a bipolar NPN phototransistor (the most common variant), the important region for pair generation is the reverse-biased collector-base interface. Photocurrent generated here acts as current injected into the base of an ordinary transistor, and permits a larger current to pass from the collector to the emitter.

The behavior of a phototransistor can be visualized as being similar to a photodiode controlling an ordinary bipolar transistor, as shown in [Figure 22-3](#).

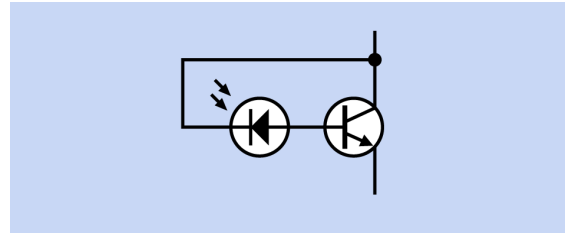


Figure 22-3 A phototransistor is functionally similar to a photodiode controlling an ordinary transistor.

Variants

While surface-mount variants are very widely available, through-hole packaging also remains common. When encapsulated like an LED indicator, a phototransistor gathers light from a relatively narrow angle. Variants with a flat surface are sensitive to light from almost any direction in front of the component.

Optional Base Connection

The base of a phototransistor is usually not accessible. However, some variants provide a base connection (or gate connection, in an FET) in addition to the collector and emitter (or the source and drain, in an FET). This third connection allows the application of a bias current, which can prevent low light levels from triggering the transistor.

Photodarlington

A photodarlington is a pair of bipolar transistors, the first being sensitive to light while the second acts as an amplifier for the first. This configuration is very similar to that of a [Darlington transistor](#) (described in Volume 1). The two-stage design makes it more sensitive to light than a regular phototransistor, but makes the response slower and less linear.

PhotoFET

A field-effect phototransistor is sometimes identified as a [photoFET](#). They are relatively uncommon as separate components, but are used in optocouplers, because they have some interesting properties:

- Provided the applied voltage is low enough (less than 0.1V) the photoFET works as a controllable resistance—in contrast to bipolar transistors, where the current is controlled, and is relatively independent of the applied voltage.
- The FET transistor is symmetrical, functioning similarly for signals of either polarity. This makes an FET optocoupler suitable for AC signals.

Values

Abbreviations found in datasheets are included in the list below, with values in parentheses for an Osram SFH300FA infrared photodiode, which resembles a 5mm through-hole LED. It has a peak sensitivity of 880nm and appears black to the human eye, being opaque to wavelengths of light shorter than 700nm, the red end of the visible spectrum.

In [Figure 22-4](#) the SFH300FA is shown beside the SFH300, which has the same peak sensitivity of 880nm but is encapsulated in an untinted module, allowing a sensitivity that tapers gradually to 450nm, in the green part of the visible spectrum. With the exception of their spectral range, the two phototransistors have identical specifications.

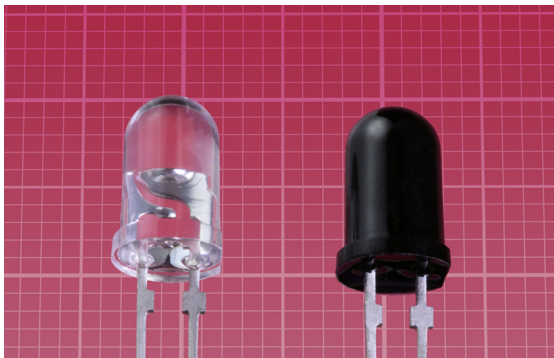


Figure 22-4 Osram infrared phototransistors SFH300 (left) and SFH300FA (right). The background grid is in millimeters.

- Maximum collector-emitter voltage: V_{CE} (35V)
- Maximum collector current: I_C (50mA)
- Maximum power dissipation: P_{TOT} (200mW)
- The *half angle* is measured from the axis of the photodiode to the angle at which sensitivity has dropped by 50% (plus or minus 25 degrees)

The angle of sensitivity depends on the geometry of the package. For phototransistors that resemble an LED indicator, with a rounded end that acts as a lens, typical values are plus-or-minus 20 degrees.

- Dark current (when the phototransistor receives no incident light): I_{CE0} (1nA)
- Wavelength of maximum sensitivity: $\lambda_{S\ MAX}$ (880nm)

Infrared phototransistors exist with a variety of peak wavelengths. They are designed to function in conjunction with an LED that has a matching wavelength.

- Response speed is the rise and fall time of photocurrent: t_r and t_f (10 μ s)

Behavior Compared to Other Light Sensors

An extended list of comparisons between a photoresistor and a phototransistor will be found in the entry on **photoresistors**. See [“Comparisons with a Phototransistor”](#).

A photodiode has a close-to-linear electrical response over a much wider range of intensities of light than a phototransistor. Consequently, photodiodes tend to be the component of choice in applications where measurement of light must be wide-ranging and precise.

Photodiodes can pass less current than a phototransistor, but also tend to draw less current,

making them appropriate in battery-powered devices that must draw as little current as possible.

Rise and fall speeds are important for high-speed measurement, signaling, or data transfer. The rise and fall time of a typical phototransistor can be 1,000 times slower than that of a photodiode. See “Values” in [Chapter 21](#) for a comparison. Note also that the dark current of a phototransistor is much higher than that for a photodiode.

The ability of a phototransistor to sink 20mA to 50mA at its output is useful where it will be connected to a component that has relatively low impedance. For instance, a phototransistor can drive a piezoelectric audio transducer directly, or an LED indicator.

Unlike a photodiode, a phototransistor is a solid-state switch. Its saturation voltage (listed in datasheets as $V_{SE(SAT)}$ as described above) is the voltage drop between collector and emitter, and seldom exceeds 0.5V.

Binning

Small variations that occur during the fabrication process can cause inconsistency in the performance of phototransistors that share the same part number. To provide a more consistent response, manufacturers use [binning](#), meaning that units sharing the same bin number are likely to share a tighter tolerance. (The same concept is used to minimize variations in **LED area lighting**, described in Volume 1.)

Datasheets will provide information on the availability and meaning of bin numbers, if available.

Bins with higher photocurrents typically have longer response times.

How to Use It

Most phototransistors are bipolar devices with an [open collector output](#). That is, the collector of the transistor is accessible from one of the two

leads, making it “open” to being used. See [Figure A-4](#) for general instructions on using an open-collector output. A summary relating to phototransistors is included here.

The schematic in [Figure 22-5](#) shows the basic concept. The resistor is referred to as a [pullup resistor](#). When the phototransistor is receiving very little light, its effective resistance is high. Consequently almost all the current flowing through the pullup resistor will go to any device attached to the output, and the output voltage will seem to be “high.”

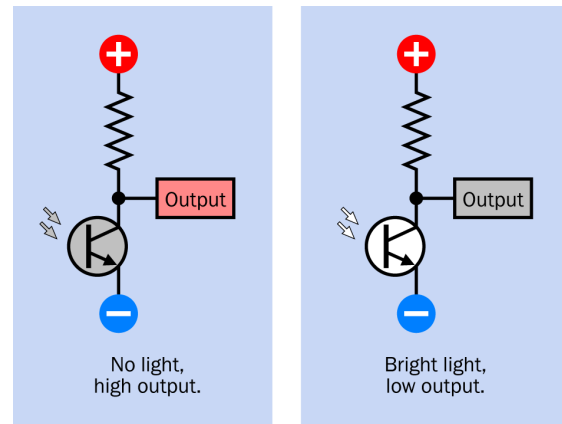


Figure 22-5 How an open-collector output works.

If the phototransistor is exposed to a significant light source, the effective resistance between the collector and emitter drops dramatically, and the phototransistor will sink current to ground. Consequently, the output will seem to be “low.”

The pullup resistor is necessary between the power source and the collector pin to protect the phototransistor from sinking excessive current when it is exposed to light. The ideal value of the resistor will depend partly on the impedance of any device attached to the output.

In this scenario, exposure to light causes low output whereas darkness causes high output. What if we wish to have it the other way around?

The protective resistor can be moved to the emitter pin, where it becomes a pull-down resistor. It will still protect the phototransistor from passing excessive current, so long as the output is connected to a high impedance. The output is taken from the emitter pin, and will transition from low to high when the component is exposed to light. This is illustrated in Figure 22-6.

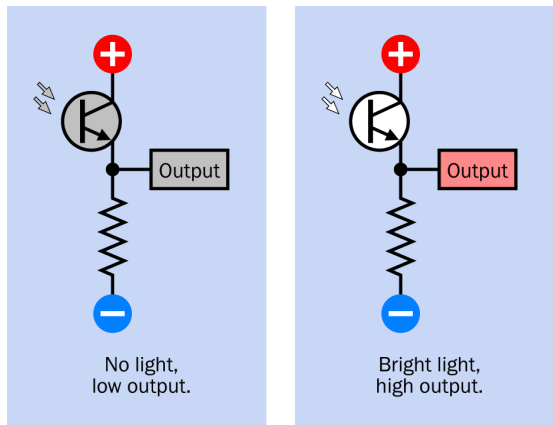


Figure 22-6 Moving the resistor and taking an output from the emitter pin inverts the behavior of a phototransistor. Any device attached to the output must have a relatively high impedance to protect the phototransistor from excessive current.

Output Calculation

Using an open-collector output, the photocurrent is almost independent of the applied voltage V_{CE} , provided that the voltage is higher than the saturation voltage $V_{CE(SAT)}$, which is typically between 0.4V and 0.5V.

If the pullup resistor has value R , the voltage across it is

$$U = R * I_p$$

where I_p is the photocurrent passed by the phototransistor.

When choosing R , one should consider the range of currents expected in the light condi-

tions that will be measured, and the voltage range suitable for the next stage of the circuit. 10K is a reasonable starting point (e.g., when measuring light intensity with a microcontroller's analog input). The resistor value can be reduced from there, if necessary.

The value of the pullup resistor must also be chosen to restrict current within the limits specified by the phototransistor datasheet. A value for the resistor guaranteed to be safe is

$$R = V / I_{MAX}$$

where V is the supply voltage and I_{MAX} is the maximum allowed current. With this value, the resistor limits the current to the highest allowed value, even if the phototransistor is brightly illuminated and assumed to conduct perfectly.

When $V = 5V$ and $I_{MAX} = 15mA$, R should be at least 330 ohms.

What Can Go Wrong

Visual Classification Errors

Phototransistors can be visually similar to LEDs and photodiodes. They are easily confused, as neither type of component is usually marked with any identification number. The entry for **photodiodes** describes a system for distinguishing the three types of components. See "What Can Go Wrong" in Chapter 21.

Output Out of Range

The output voltage from a phototransistor will depend on the intensity of incident light, the value of any pullup resistor that is used, and the supply voltage. While a circuit is being developed, the light range may seem predictable, but in actual use the output range may not fall within expectations.

NTC thermistor

23

PTC thermistors, in which the resistance increases as the temperature increases, have a separate entry. See Chapter 24.

A *resistance temperature detector* or **RTD** has a resistance that increases as its temperature increases, but it is not usually classified as a thermistor, because its sensing element is fabricated differently. Its entry will be found at Chapter 26.

Semiconductor temperature sensors and **thermocouples** each have their own entries.

Infrared temperature sensors and **passive infrared** motion sensors have their own entries. They are *noncontact* temperature sensors that respond to infrared radiation.

OTHER RELATED COMPONENTS

- **PTC thermistor** (see Chapter 24)
- **infrared temperature** sensor (see Chapter 28)
- **passive infrared** motion sensor (see Chapter 4)
- **semiconductor** temperature sensor (see Chapter 27)
- **thermocouple** (see Chapter 25)
- **RTD** (resistance temperature detector) (see Chapter 26)

What It Does

An NTC thermistor is the most common type of discrete-component temperature sensor, and is usually the most affordable. Its resistance diminishes as its temperature increases. This behavior is referred to as a *negative temperature coefficient*, which is the source of the acronym NTC.

This is a simple, passive component that is not polarized. It requires no separate power supply, but an external device must pass a small AC or DC current through it to determine its resistance. This is known as an *excitation current*.

Schematic Symbols

Schematic symbols for a thermistor are shown in Figure 23-1. Those in the top row may still be found in the United States, but are being replaced by the European variants in the second row. The addition of $-t^{\circ}$ to the symbol indicates an NTC type of thermistor, while $+t^{\circ}$ indicates that it is the PTC type, with a positive coefficient (see Chapter 24). If no indication is shown, the thermistor is likely to be the NTC type.

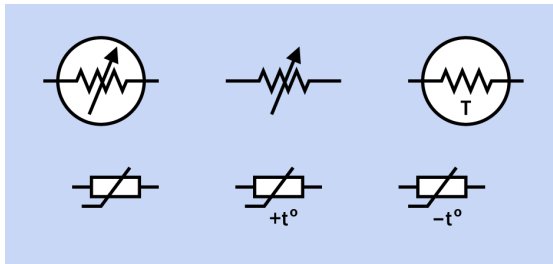


Figure 23-1 Schematic symbols representing a thermistor. Letter t preceded by a plus or minus sign indicates whether the thermistor is the PTC or NTC type, respectively.

Applications

Thermistors monitor temperature in air-conditioning systems, clothes washers, refrigerators, pool and spa controls, dishwashers, toasters, and other domestic devices. They are used in laser printers, 3D printers, industrial process controls, and medical equipment.

As many as 20 thermistors may be found in a modern automobile, measuring temperature in locations ranging from the transmission to the ambient air in the passenger compartment.

Comparison of Temperature Sensors

In this Encyclopedia, contact temperature sensors, which measure temperature by making contact with the source, are divided into five main categories, each of which has a separate entry. For convenience, these categories are listed in a comparative summary at the end of this entry. See ["Addendum: Comparison of Temperature Sensors"](#).

How an NTC Thermistor Works

Although the term *thermistor* suggests that it is a thermally sensitive resistor, in fact an NTC thermistor is a semiconductor.

Some metal oxides, such as ferric oxide or nickel oxide, become n-type semiconductors

when they are treated with dopants. The exact mix is a proprietary secret of each manufacturer. Raising the temperature of this kind of material increases the number of charge carriers in it, promoting electron mobility and thus lowering its effective resistance.

To create a thermistor, the metal oxide mix is heated until it melts and turns into a ceramic. Typically a thin sheet is cut into small pieces for individual sensors. After two leads are connected, the assembly is dipped into epoxy or encapsulated in glass. The most common packages consist of a glass bead, surface-mount chip, or ceramic disc.

Figure 23-2 shows three NTC thermistors. At left is a Murata NXFT15XH103FA2B100 approximately 1mm in diameter, with a reference resistance of 10K and a tolerance of plus-or-minus 1%. At center is a Vishay NTCA-LUG03A103GC rated 10K at 2%, fitted with a mini-lug. At right is a TDK B57164K153K rated 15K and 3%.

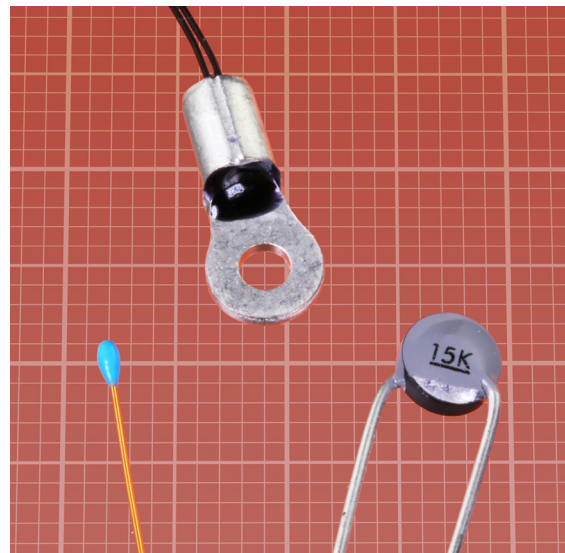


Figure 23-2 Sample NTC thermistors. See text for details. The background grid is in millimeters.

Output Conversion for Temperature Sensing

Ideally, the electrical behavior of a temperature sensor should be a linear function of temperature. Thermistors fail in this respect, as their resistance is an approximately inverse exponential function. This is illustrated in Figure 23-3, where the measured resistance of a thermistor rated for 5K at 25 degrees is plotted against temperatures from 0 degrees to 120 degrees Celsius.

- In many datasheets, graphs of this kind may appear flatter because they are customarily plotted against a vertical logarithmic scale.

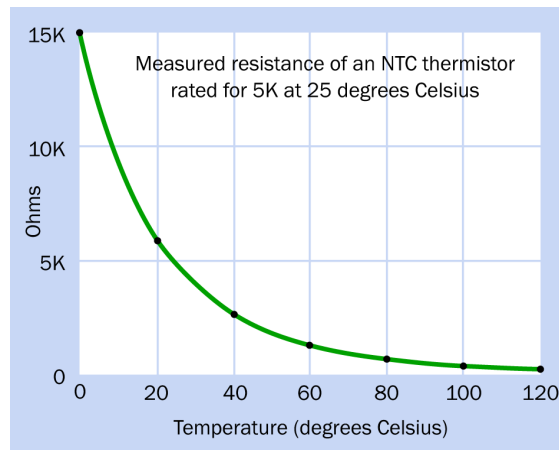


Figure 23-3 Resistance of a thermistor from 0 to 120 degrees Celsius.

To monitor the resistance of a thermistor, it can be placed in a simple voltage divider as shown in Figure 23-4, where the fluctuating resistance of the component creates a fluctuating voltage at point A.

- The voltage can be used as an input to a microcontroller that contains an analog-to-digital converter. Alternatively it can be connected directly to a solid-state relay, or amplified with an op-amp, or can be passed through a

comparator to create an adjustable switching threshold.

Although this circuit is a voltage divider, it is also known as a *half bridge*, as it is half of a Wheatstone bridge.

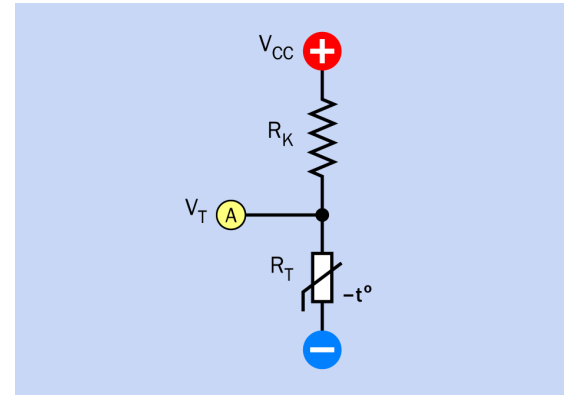


Figure 23-4 A half-bridge circuit for determining the resistance of a thermistor.

If V_{CC} is the supply voltage, V_T is the measured voltage at point A, R_T is the resistance of the thermistor, and R_K is the constant value of the series resistor, the basic formula for a voltage divider looks like this:

$$V_T = V_{CC} * (R_T / (R_T + R_K))$$

By transposing terms, a formula can be derived to obtain a value for R_T from the measured voltage and the value of R_K :

$$R_T = (R_K * V_T) / (V_{CC} - V_T)$$

Choosing a Series Resistor

The value for R_K in the formula should be chosen to provide a reasonably wide response over the range of temperatures for which the thermistor is likely to be used. To calculate R_K , another formula must be applied. If R_{MIN} is the resistance of the thermistor at the lowest likely temperature, and R_{MAX} is its resistance at the highest likely temperature:

$$R_k = \sqrt{R_{min} * R_{max}}$$

(This is the same formula as suggested in [Figure 20-5](#) to find the value of a series resistor for use with a photoresistor.)

Wheatstone Bridge Circuit

The half-bridge circuit has the disadvantage that it does not compensate for nonlinearity of a thermistor. Voltage values will change rapidly at the low end of the temperature range, but will change more slowly at the high end, requiring an analog-to-digital converter with a high degree of accuracy to distinguish one voltage value from the next.

A full Wheatstone bridge circuit has a nonlinear output that compensates, somewhat, for the inverse nonlinearity of the thermistor. Referring to the circuit shown in [Figure 23-5](#), the three resistors R_K are chosen using the formula above.

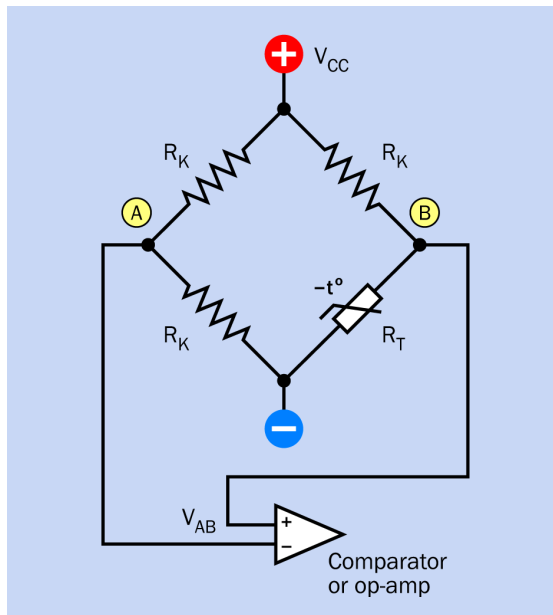


Figure 23-5 A thermistor may be placed in a full Wheatstone bridge circuit. Outputs A and B are often connected with the two inputs of an op-amp or comparator.

A standard formula provides the relationship between R_T , the resistance of the thermocouple; V_{CC} , the supply voltage; R_K , the fixed resis-

tances; and V_{AB} , the output voltage measured between points A and B:

$$V_{AB} = (V_{CC} / 2) * (R_T - R_K) / (R_T + R_K)$$

From this formula, a version can be derived to calculate R_T by measuring the output voltage, V_A :

$$R_T = R_K * (V_{CC} + (2 * V_{AB})) / (V_{CC} - (2 * V_{AB}))$$

- The polarity of V_{AB} is reversible, depending on whether R_T is greater or less than R_K . To accommodate this, A and B can be connected to the two inputs of a comparator or op-amp.

Deriving the Temperature Value

After the resistance of the thermistor has been calculated, it can be converted to a temperature value. The datasheet for a thermistor will usually provide a table showing temperature values tabulated against resistance values, so that a lookup table can be created in a microcontroller program.

Alternatively, a datasheet usually includes constants that can be inserted in a resistance-to-temperature conversion equation, but this is nontrivial and requires natural logarithms, which may not be available in a language implemented on a particular microcontroller.

Inrush Current Limiter

NTC thermistors with appropriate characteristics can be used to limit the inrush of current that tends to occur when a circuit is switched on and large capacitors in the power supply charge very quickly.

An *inrush current limiter* is also known as a *surge limiter*, or may be referred to by its acronym, *ICL*. It is an NTC thermistor whose initial resistance diminishes rapidly as its temperature increases.

While NTC thermistors are the type most often used for inrush limiting, PTC thermistors can

serve this purpose if wired differently. See “PTC Inrush Current Limiting”. The remainder of this discussion refers only to NTC current limiters.

A suitable NTC thermistor can be placed as shown in the simplified schematic in Figure 23-6, where a rectified AC source is connected with a DC-to-DC converter, and a large smoothing capacitor is used. Initially the thermistor has resistance that is sufficient to limit current and generate heat. But the rise in temperature causes the resistance of the thermistor to fall. Eventually it reaches a steady state where it remains sufficiently warm to maintain a low resistance that imposes a negligible load on the circuit.

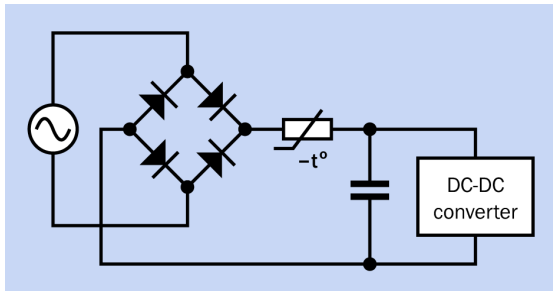


Figure 23-6 Placement of an NTC thermistor that is designed for inrush current limiting.

In thermistors that are used for temperature measurement, self-heating is an undesirable attribute. By contrast, an inrush current limiter depends on self-heating to perform its function.

The TDK B57237S509M inrush limiter, shown in Figure 23-7, is rated for 5A and has an initial resistance of 5 ohms at 25 degrees Celsius while not passing current. When tested with a 2,800 μ F capacitor at 110VAC, its resistance drops to a minimum of 0.125 ohms at 5A. The relationship between current and resistance is shown in Figure 23-8.

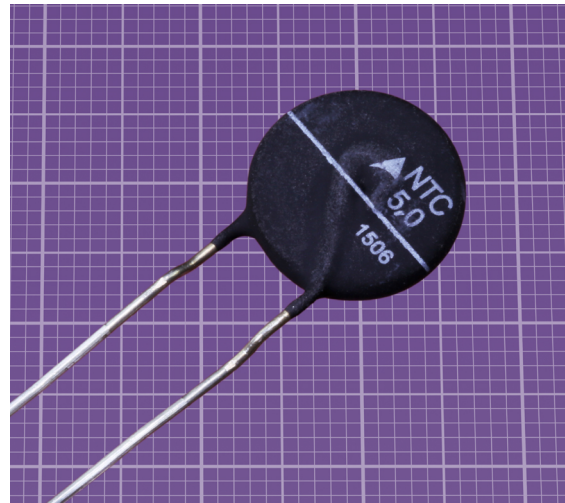


Figure 23-7 A TDK B57237S509M NTC thermistor designed as an inrush current limiter, rated 5 ohms at 5 amps. The background grid is in millimeters.

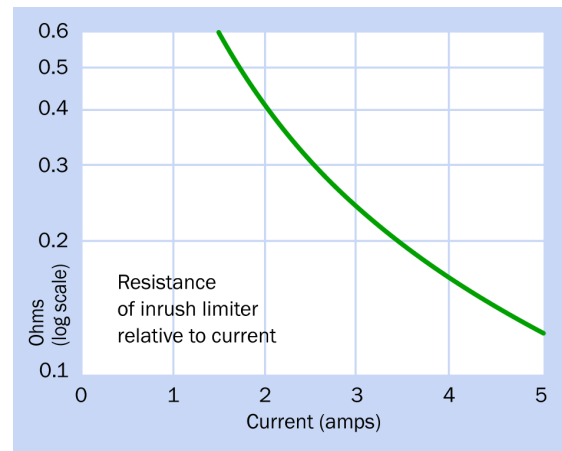


Figure 23-8 This graph shows the relationship between resistance and current in a TDK B57237S509M inrush limiter.

Restart

If a protected device is switched off momentarily and is then switched on again, the thermistor cannot provide protection, as it has not had time to cool down and regain its resistance. However, during the 30 seconds to 2 minutes required for heat in the thermistor to dissipate, smoothing capacitors are unlikely to lose much of their charge. Therefore, if the device is restarted, an inrush of current should not occur.

Thermistor Values

Datasheets for thermistors may be more complex and cryptic than for many components.

When examining a datasheet, first check to see if the thermistor is described as being suitable for temperature measurement or inrush current limiting. A component designed for temperature measurement will not survive inrush current, while one designed for inrush current limiting will have a very low resistance making it unsuitable for temperature measurement.

Time and Temperature

In most datasheets, lowercase letter t is used for values relating to time, whereas an uppercase T is used for values relating to temperature. Unfortunately, T may also be used as an abbreviation for “thermistor.”

Resistance and Response

Letter R often means resistance, but may indicate response time, depending on the context in which it is used. For example, R_T is the resistance of a thermistor, and t_R is a response time.

Kilohms and Kelvin

Letter K may be used to represent temperature in degrees Kelvin, 0 degrees on the Celsius scale being approximately 273 degrees Kelvin. However, letter K is also used to represent thousands of ohms, sometimes in the same datasheet. In both instances, K is capitalized.

Reference Temperature

This is the temperature at which many attributes of the component are measured, such as its temperature coefficient and resistance. Usually the reference temperature is 25 degrees Celsius, but in some cases it may be 0 degrees, and other values are occasionally used. The term is abbreviated T_{REF} .

Reference Resistance

The reference resistance for a thermistor (sometimes described as its nominal resistance) may

be referred to as R_R , and is the resistance at the reference temperature. It may be referred to as the “R value,” but in thumbnail product descriptions it can be cited simply as “Resistance.”

In datasheets, R_{25} or R_{25} is the resistance at 25 degrees Celsius. If this is the reference temperature, R_R and R_{25} will be the same.

Dissipation Constant

DC is the *power dissipation constant*, a ratio normally expressed as milliwatts per degree Celsius (written as mW/°C). This is a measurement of how much thermal power the thermistor can transfer to the environment for a 1 degree increase in temperature.

Temperature Coefficient

TC may be used as an acronym for the *temperature coefficient*, which represents the sensitivity of the thermistor. (Sometimes TCR is used instead of TC, the letter R denoting resistance. The two acronyms both mean the same thing.) The value is the percentage change in resistance for each change in temperature of 1 degree Celsius. Thus, if the resistance of a thermistor drops from 800 ohms to 768 ohms when the temperature increases from 28 to 29 degrees Celsius, $TC = -4\%$. For NTC thermistors, which have a resistance that decreases when temperature increases, the temperature coefficient is negative. However, the minus sign may be omitted.

The coefficient may be expressed in parts per million (abbreviated ppm) instead of as a percentage. To convert parts per million to a percentage, divide by 10,000. Thus, a figure of 50,000ppm is equivalent to 5%.

Thermal Time Constant

Unfortunately TC is also used to represent the thermal *time constant*. If T_D is the temperature difference between the thermistor's initial temperature and a new, higher ambient temperature in which it finds itself, TC is the time it takes for the thermistor to add 63.2% of T_D to its cur-

rent temperature. TC is expressed in seconds, and is defined without power being applied to the thermistor. A low thermal time constant is characteristic of a physically small thermistor that acquires heat rapidly. (TC is very similar to the concept of a time constant for a capacitor acquiring charge. See the entry on **capacitors** in Volume 1.)

Tolerance

The *tolerance* of a thermistor is a measure of its accuracy, usually at 25 degrees Celsius, unless a range of temperatures is stated. A thermistor rated for 5K, with a tolerance of plus-or-minus 1% at 25 degrees, may be found to have an actual resistance ranging from 4,950 to 5,050 ohms at that temperature. Some thermistors have a tolerance of plus-or-minus 20%. A tolerance better than plus-or-minus 1% is relatively rare.

Temperature Range

The *working temperature range* of any thermistor that uses silicon dioxide is usually between about -50 and +150 degrees Celsius (slightly wider for versions encapsulated in glass, and slightly narrower if accuracy is important).

Switching Current

For a thermistor with a nonlinear response, the switching current is the approximate current that forces a sharp transition in resistance. It is represented by I_S .

Power Limitations

Operating current is the maximum current recommended to avoid self-heating. The *power rating* is the maximum allowed power (usually 100mW to 200mW).

Interchangeability

To measure temperature reliably, two thermistors of the same type, from the same manufacturer, should display the same characteristics. This is known as *interchangeability*. A value of plus-or-minus 0.2 degrees Celsius is common

for a modern thermistor, but is not often mentioned in a datasheet.

What Can Go Wrong

Self-Heating

Self-heating can affect the accuracy of an NTC thermistor that is used for temperature measurement. To get accurate temperature readings, keep the current as small as possible. When the resistance of a thermistor is at the high end of its range, brief pulses of current can be used.

Heat Dissipation

Where a thermistor is used for inrush current limiting, it will create some heat during the whole time that a device is switched on. If insufficient air space is allowed between the thermistor and other components, they may be affected.

Lack of Heat

An NTC thermistor will sometimes fail as an inrush current limiter. In very cold climates, it may never become warm enough for its resistance to drop to an acceptable level. Conversely, in a very hot location (such as close proximity to a hot-water pump) it may not get cool enough to provide adequate initial protection.

Addendum: Comparison of Temperature Sensors

A chart illustrating the five main types of contact sensors, and their variants, is shown in [Figure 23-9](#).

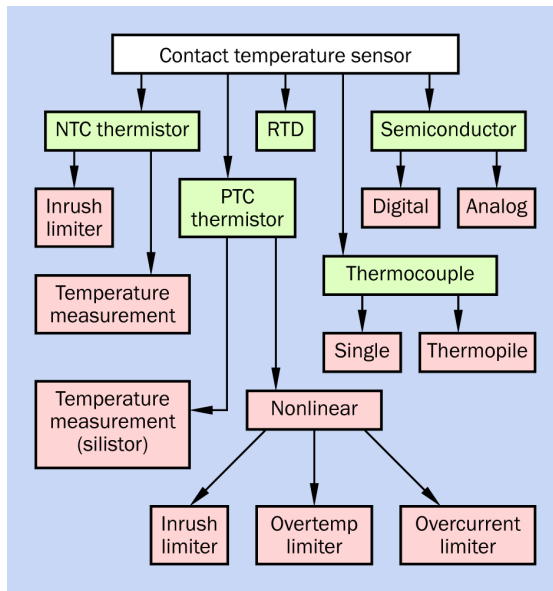


Figure 23-9 Five types of contact temperature sensors (green boxes) and the variants (red).

NTC Thermistor

The electrical resistance of an **NTC thermistor** diminishes as its temperature increases. Thus, it has a *negative temperature coefficient*, which is the source of the acronym NTC.

An NTC thermistor is traditionally used where low cost and simplicity are desirable and a relatively limited temperature range is acceptable (often -50 to +150 degrees Celsius). It has the advantage of familiarity, having existed in its present form for many decades. It remains the lowest-cost option among the various types of temperature sensors, and can be connected directly with an external device such as a solid-state relay, in which case no microcontroller is necessary.

PTC Thermistor

The sensing element for a positive-coefficient thermistor is a polycrystalline compound that increases in resistance very rapidly above a threshold temperature. This makes it suitable for blocking a high current to prevent circuit overload.

A *silicon temperature sensor*, sometimes called a *silistor*, can be considered as a PTC thermistor, in that it is a resistive component with a positive temperature coefficient. Its sensing element is etched into silicon.

PTC thermistors are passive, nonpolarized components with two leads or solder pads. For more information about them, see [Chapter 24](#).

Thermocouple

This sensor consists of two wires made from different metals, joined at one end. The differing thermoelectric properties of the wires creates a very small voltage between their free ends. Thermocouples have the widest range of any contact sensor. They are simple, robust, and free from self-heating effects, as they consume no power. Their response is rapid, but very non-linear, and their sensitivity is limited. They are used in industry and in laboratories, often plugged into a panel meter that combines a digital temperature display with hardware to decode the signal from the type of thermocouple being used.

For more information about thermocouples, see [Chapter 25](#).

Resistance Temperature Detector

Often known by its acronym **RTD**, and sometimes referred to as a *Resistive Temperature Device*, it commonly uses a sensing element fabricated from pure platinum, nickel, or copper. The element may consist of wire wound around a core, or a very thin film deposited on an insulating substrate.

An RTD has a positive temperature coefficient, as its resistance increases while its temperature increases. It is very accurate and stable, providing an almost linear output, especially near the center of its range. However, its sensitivity is often one-tenth of that of an NTC thermistor.

Like a thermistor or a thermocouple, an RTD is a passive device, able to operate at a wide range

of voltages and requiring no power supply. It is nonpolarized, with two leads or solder pads.

For more information about resistance temperature detectors, see [Chapter 26](#).

Semiconductor Temperature Sensor

This is a chip-based sensor that requires no additional components to linearize its output, as linearization is performed in the chip.

The temperature range is similar to that of an NTC thermistor, but the output is a variable voltage with a positive temperature coefficient of about 20mV per degree Celsius, supplied by a built-in op-amp. Response time is 4 to 60 seconds.

This type of sensor requires a power supply, typically of 5VDC or less. It does not have to be calibrated before use, as it is trimmed during the production process to achieve accuracy that can be better than that of a thermistor. Manufacturers may claim plus-or-minus 0.15 degrees over the whole temperature range, which is usually -50 degrees Celsius to +150 degrees, but may be less for variants in which accuracy is more important.

The linear analog output is very convenient for use with a microcontroller that has an analog-to-digital converter, and the relatively low cost makes this type of sensor increasingly competitive with thermistors.

An analog-to-digital converter may be included on the sensor chip, in which case it is often described as a *digital temperature sensor* or *digital thermometer*, with an output in degrees Celsius (or, sometimes, Fahrenheit) accessible via I2C or SPI bus. For additional details about protocols such as I2C and SPI, see [Appendix A](#).

A *digital thermostat* or *thermostatic switch* is a semiconductor temperature sensor with a binary output that transitions from logic-high to logic-low (or vice-versa) if the temperature goes above a maximum or below a minimum level. The level can be programmed into the chip.

Semiconductor temperature sensors are identified with a variety of other names. For more information, see [Chapter 27](#).

PTC thermistor

24

A *silistor*, or *silicon-based thermistor*, is included in this entry as a form of **PTC thermistor**.

A *resettable fuse* is not quite the same as a PTC thermistor. For more information, see the entry on **fuses** in Volume 1.

NTC thermistors, in which the resistance decreases as the temperature increases, have a separate entry. See Chapter 23.

A *resistance temperature detector* or **RTD** has a resistance that increases as its temperature increases, but it is not usually classified as a thermistor, because its sensing element is fabricated differently. Its entry will be found at Chapter 26.

Infrared temperature sensors, **semiconductor temperature** sensors, and **thermocouples** each have their own entries.

OTHER RELATED COMPONENTS

- **infrared temperature** sensor (see [Chapter 28](#))
- **semiconductor** temperature sensor (see [Chapter 27](#))
- **thermocouple** (see [Chapter 25](#))
- **NTC thermistor** (see [Chapter 23](#))
- **RTD** (resistance temperature detector) (see [Chapter 26](#))

What It Does

The electrical resistance of a PTC thermistor increases as its temperature increases. Variants can measure temperature or can protect circuits by detecting excessive heat or current.

Because a PTC thermistor is a resistive sensor, it has no polarity. Current may flow through it in either direction, or AC may be used.

Schematic Symbols

The schematic symbol for a PTC thermistor is very similar to the symbol for an NTC thermistor. See [Figure 23-1](#).

Comparison of Temperature Sensors

In this Encyclopedia, contact temperature sensors are divided into five main categories, each of which has a separate entry. For convenience, a comparative summary is included in the entry for

NTC thermistors. See “Addendum: Comparison of Temperature Sensors”. Also see Figure 23-9.

PTC Overview

PTC thermistors can be subdivided into two groups:

- **Linear**, with a chip-sized silicon-based sensing element. They are sometimes referred to as **sili-stors**. The component has a very linear response and is used for temperature measurement. It may be connected directly to a microcontroller.
- **Nonlinear**, mostly using a sensing element containing barium titanate in a polycrystalline compound that increases in resistance very sharply above a threshold temperature. This type of sensor may be described as a **switching thermistor**, because its nonlinear output can activate a switching device.

The sensing elements in positive-coefficient thermistors are different in principle from the element in an NTC thermistor.

Nonlinear thermistors are used in two different ways:

Externally heated

The thermistor responds to ambient heat or to the temperature of a device to which it is attached. It can be used to protect a circuit or a motor from overheating. Current through the thermistor is minimized to avoid self-heating.

Internally heated

The thermistor responds to its own temperature caused by current passing through it. It can activate a warning signal or shut down equipment in the event of a short circuit. It can also control current for starting a motor or a fluorescent tube, and is sometimes used as a source to create localized heat.

Silistor for Temperature Measurement

A silicon-based PTC thermistor, sometimes known as a **silistor**, provides a highly desirable, almost linear relationship between temperature and resistance. A popular example is the KTY81 series from NXP, a sample of which is shown in Figure 24-1.

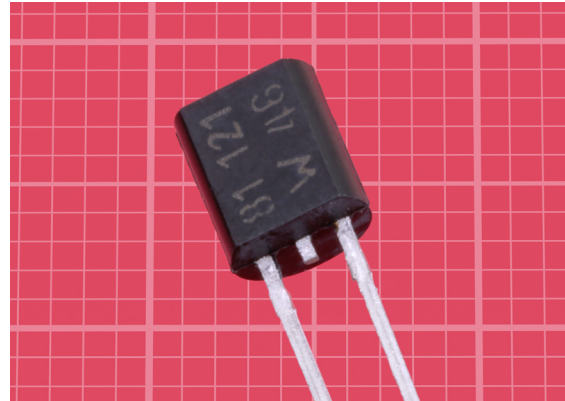


Figure 24-1 A KTY81 thermistor from NXP. The background grid is in millimeters. Note the amputated center lead.

The response of this thermistor is shown in Figure 24-2.

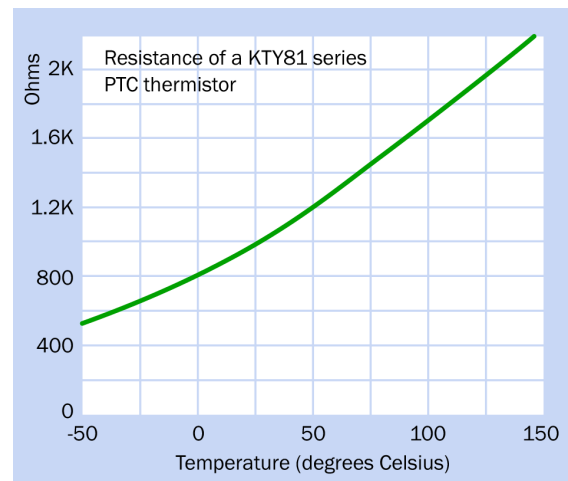


Figure 24-2 Resistance of the KTY81 thermistor in response to temperature.

Note that this graph is plotted with a linear vertical scale, unlike the performance curves for many thermistors that are plotted with a log scale. The log scale tends to make a response curve look flatter.

The sensor is a silicon chip designed on the “spreading resistance principle,” in which current fans out from a metal contact through a thin layer of silicon to a metallized bottom plane. This effect progresses less actively as the temperature increases. Although the result is partly dependent on polarity, a second metal contact is biased in the opposite direction, and when the two active regions of the chip are wired in series, the result is a component that has no polarity.

- The almost-linear output of this type of sensor makes it easy to use with a microcontroller that has a built in analog-to-digital converter.

Tolerance ranges from plus-or-minus 1% to 5%, depending on the temperature. Variants have a typical reference resistance of 1K or 2K. The temperature coefficient is commonly about 1%, which is considerably lower than that of a typical NTC thermistor, where 4% is common.

- Guidance on reading thermistor data-sheets will be found in the entry describing NTC thermistors. See “[Thermistor Values](#)”.

For correct operation, a typical silistor requires a current ranging from around 0.1mA to 1mA.

- The lower sensitivity and slightly higher price of a PTC temperature-measurement thermistor, compared with an NTC thermistor, may explain why the NTC type seems to remain more popular, with many more variants available. In addition, the NTC type is much more tolerant of variations in current.

- Silistors continue to find some automotive applications, measuring oil temperature, transmission temperature, and climate control, among other parameters.

As a simple strategy to determine its resistance, a series resistor can be used with a PTC sensor to create a voltage divider. The circuit is identical to that used for NTC thermistors. See “[Output Conversion for Temperature Sensing](#)”.

RTDs

A *resistance temperature detector* or **RTD** is sometimes classified as a PTC thermistor. However, it has a different type of pure-metal sensing element, much lower sensitivity, and is discussed in a separate section of this Encyclopedia. See [Chapter 26](#).

Nonlinear PTC Thermistors

Over-Temperature Protection

This type of nonlinear thermistor is externally heated, but has a switching function. If it is incorporated among other components on a circuit board, its output can be used to activate a warning signal, or can trigger a relay to shut down the circuit until the temperature subsides. This is of special interest for battery chargers where excessive heat can often be a problem, but is also useful in electronic devices generally.

To avoid the possibility of self-heating, current passing through the thermistor must be minimized to a few milliamps.

Some thermistors in the Vishay PTCSL series will make a transition at a temperature as low as 70 degrees Celsius. Others will be triggered by temperatures above 100 degrees. A typical response curve is shown in [Figure 24-3](#), where resistance rises from 100 ohms at 25 degrees to around 1K at the transitional reference temperature of 90 degrees, and reaches at least 4K at 105 degrees.

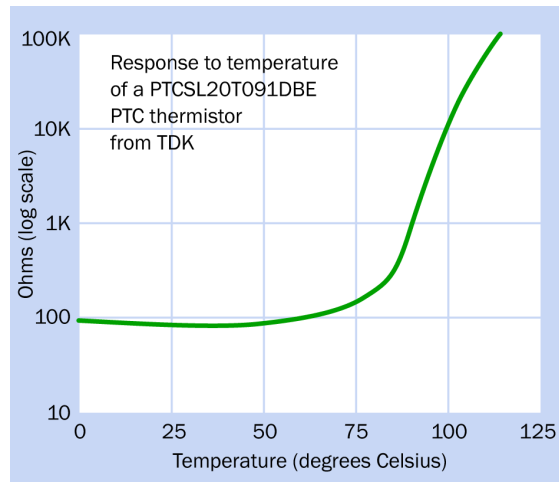


Figure 24-3 The relationship of resistance to heat in an over-temperature protection thermistor.

To respond to this transition, the manufacturer recommends a Wheatstone bridge circuit with its outputs connected to a comparator, as suggested for an NTC thermistor in [Figure 23-5](#). The comparator can then activate a signal or a relay.

A picture of the PTCSL20T091DBE thermistor appears in [Figure 24-4](#).

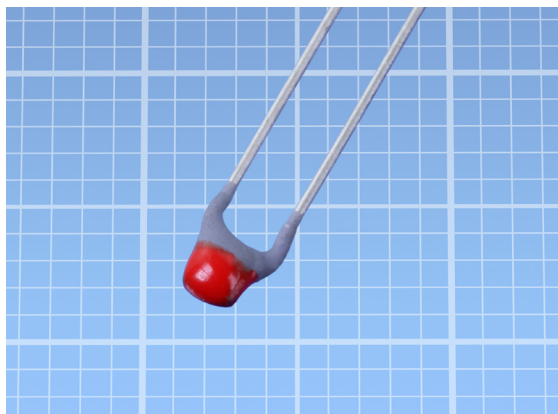


Figure 24-4 A thermistor in the PTCSL range from TDK. It is color coded using a proprietary scheme by the manufacturer to indicate a reference temperature of 90 degrees Celsius. The background grid is in millimeters.

This type of thermistor can tolerate a maximum of 30V (AC or DC).

Over-Current Protection

This type of nonlinear thermistor is a substitute for a fuse, as it responds to internal heat created by current passing through it. If the flow of current is excessive, the resistance of the thermistor increases, throttling the flow. When the over-current problem is resolved, the thermistor returns to its normal state. Whereas a fuse must be placed in a location allowing replacement, the thermistor is unharmed by its transition and does not have to be replaced.

Over-current may occur as a result of failure of other components, such as a rectifier diode or a capacitor, or can occur in situations such as a DC motor locking up.

The B598 series from TDK can tolerate voltages over 240V, AC or DC. They typically respond when currents exceed 100mA to 1A, depending on the specific component (a few fall outside that range), and many can withstand 1A to 7A. The B59810C0130A070 pictured in [Figure 24-5](#) is switched by 980mA, can tolerate as much as 7A, and has a reference resistance of 3.5 ohms, rising above 10K when excessive current causes sufficient heat.

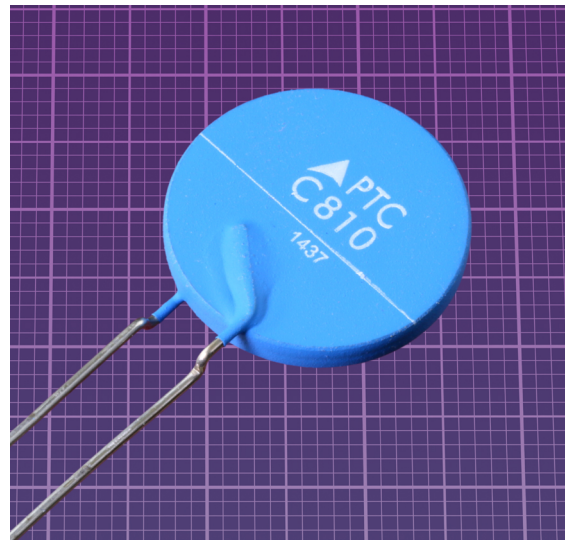


Figure 24-5 A large over-current protection PTC thermistor. The background grid is in millimeters.

An over-current thermistor of this type remains wired into the power supply for a device on a permanent basis. Its reference resistance will generate some heat, which is why this type of component is usually restricted to applications where the triggering current is below 1A.

The Murata PTGL07BD220N3B51B0 pictured in [Figure 24-6](#) provides over-current protection with a reference resistance of 22 ohms, has a trip current of 200mA, and tolerates a maximum of 1.5A.

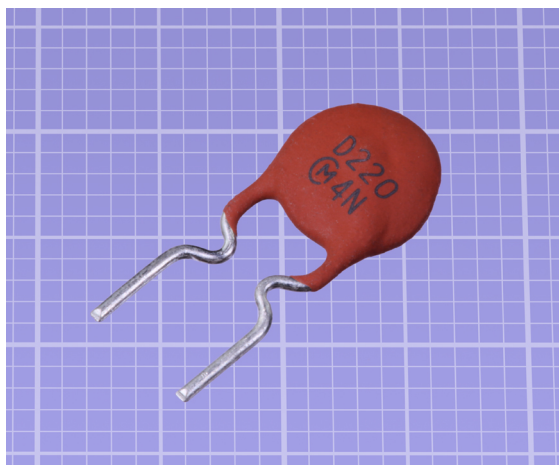


Figure 24-6 An over-current protection PTC thermistor with a trip current of 200mA. The background grid is in millimeters.

PTC Inrush Current Limiting

This nonlinear thermistor responds to internal heat caused by an inrush of current when power to a device is switched on. The inrush occurs when current flows rapidly into smoothing capacitors, charging them very rapidly. This can overload a power supply and shorten its life expectancy.

NTC thermistors are traditionally used as inrush limiters. The initially high resistance of this type of component blocks the surge in current, but as heating occurs, the resistance of the NTC thermistor drops rapidly. It remains in the circuit, imposing a relatively small load while the device functions normally. For more details of

this application, see [“Inrush Current Limiter”](#) in the entry discussing NTC thermistors.

However, an NTC thermistor used in this way will waste some power. Suppose a supply of 120VAC is being used. If the power consumption of a device is 1,000W, the current will be approximately 8A. An NTC thermistor that has a resistance of 0.2 ohms while running warm will introduce a voltage drop of approximately 1.6V, consuming about 13W. This power loss will be greater in applications where the current is even higher—for example, in an electric vehicle recharging station.

To eliminate the loss, a timed bypass relay can be added around the thermistor. The relay closes automatically after a short interval, eliminating the power loss. This is known as [active inrush current limiting](#).

However, in this arrangement, an ordinary resistor could be used instead. But in that case, why not use a PTC thermistor that has a reference (cold) resistance of 50 ohms or more? This not only limits the inrush current, but provides additional protection. If a smoothing capacitor in the circuit suffers a short circuit, or if the bypass relay fails to close, excess current passing through the PTC thermistor quickly raises its resistance, protecting the rest of the circuit.

The B5910 series of PTC thermistors from TDK is designed for inrush current limiting. They are packaged in a flame-retardant phenolic resin plastic case, as shown in [Figure 24-7](#). The B59105J0130A020 has a reference resistance of 22 ohms that rises quickly beyond 10K when the temperature exceeds 120 degrees Celsius, as shown in [Figure 24-8](#). This type of component is robust enough to withstand a complete short circuit across a 220-volt supply.



Figure 24-7 This inrush current-limiting PTC thermistor by TDK is packaged in a flame-retardant case. The background grid is in millimeters.

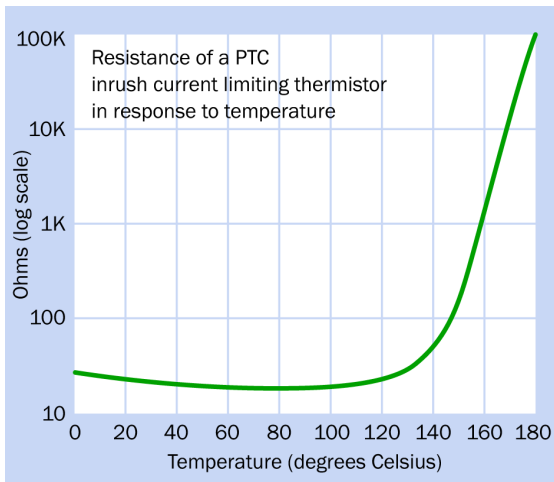


Figure 24-8 Relationship of resistance to temperature in a PTC inrush current limiting thermistor. Note that the vertical axis has a logarithmic scale.

PTC Thermistor for Starting Current

In some applications an initial inrush of current is actually necessary and desirable. An air conditioning compressor, for example, requires a

surge of high current for “torque assist” when it is starting from a rest state.

High-current PTC thermistors may be used in this kind of situation. The Vishay PTC305C series is an example. These are heavy-duty components that have a switching time of about half a second, a maximum voltage rating of 410VAC or more, and a current rating from 6 to 36 amps.

The PTC thermistor has a relatively high temperature while the motor is running, and must be allowed time to cool before a restart is possible after shutdown. A waiting time of 3 to 5 minutes is imposed by a thermostat or separate time-delay relay.

PTC Thermistor for Lighting Ballast

The starting sequence for a fluorescent lamp requires that current should flow through the cathode heater initially. The thermistor allows this by bypassing a capacitor. Within less than a second, the resistance of the thermistor rises to block current. By this time, the heater has done its job and the lamp runs from high-frequency AC.

PTC Thermistor as a Heating Element

For small applications, a heating element can be made from a PTC thermistor, using its internal resistance to create heat. It has the advantage of being self-limiting, as its resistance rises with temperature. The TDK 5906 series is an example, shown in Figure 24-9. The component is approximately 12mm in diameter, and is designed to be clamped in place, not soldered. It has automotive applications for diesel fuel preheating and spray nozzle defrosting. Residential applications include vaporizers for air fresheners.

The initial resistance is as low as 3 or 4 ohms, rising very quickly at a transition temperature ranging from 70 to 200 degrees Celsius, depending on the specific component.

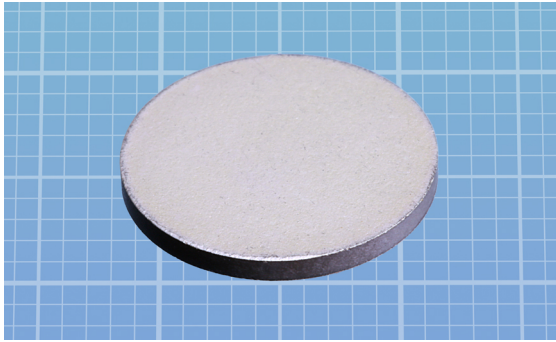


Figure 24-9 This TDK B59060A0060A010 heating element is a PTC thermistor whose resistance rises rapidly around 80 degrees Celsius. Rated for 12VDC, it is intended for automotive applications. The background grid is in millimeters.

What Can Go Wrong

Self-Heating

Self-heating may affect the accuracy of a temperature sensor. To get accurate readings, keep the current small. When the resistance of a thermistor is at the high end of its range, brief pulses of current can be used.

Heating Other Components

In cases where the self-heating of thermistors serves a useful purpose, as in surge protectors and when used for delays, the heat can damage nearby components or materials.

thermocouple

25

Because a *thermopile* is an assembly of thermocouples, it is included at the end of this entry. Other types of temperature sensors have their own entries.

OTHER RELATED COMPONENTS

- **NTC thermistor** (see [Chapter 23](#))
- **PTC thermistor** (see [Chapter 24](#))
- **semiconductor** temperature sensor (see [Chapter 27](#))
- **RTD** (resistance temperature detector) (see [Chapter 26](#))
- **infrared temperature** sensor (see [Chapter 28](#))

What It Does

A **thermocouple** measures temperature by using a pair of wires made from dissimilar metals. At one end of each wire, they are joined together, often by welding them. The differing thermoelectric characteristics of the wires generates a very small voltage between their free ends, from which the temperature of the joined ends can be derived.

No power supply is needed for a thermocouple, but the voltage that it generates is extremely small (measured not just in millivolts, but microvolts) and very nonlinear, requiring hardware and/or software to convert it to a temperature value. Laboratory equipment or integrated circuit chips are available for this purpose.

Different types of thermocouples are available to measure different temperature ranges, and each type has its own characteristics, requiring appropriate conversion.

A “raw” thermocouple looks very unimpressive, as it merely consists of two wires welded together at one end. This is illustrated in [Figure 25-1](#). The full length of the photocouple is shown in [Figure 25-2](#).

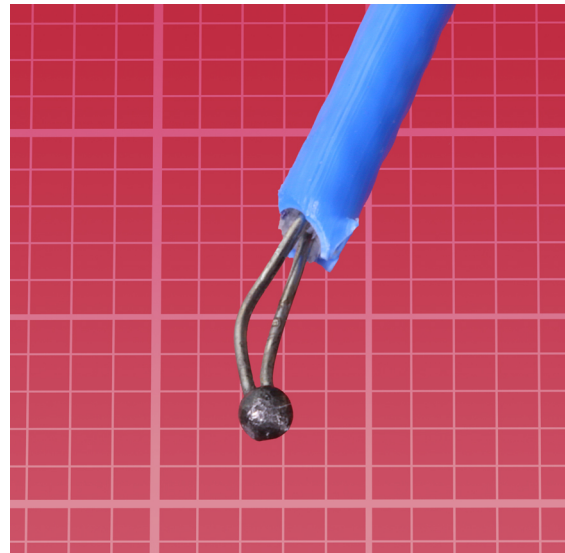


Figure 25-1 Closeup of the welded wires in a K-type thermocouple. The background grid is in millimeters.

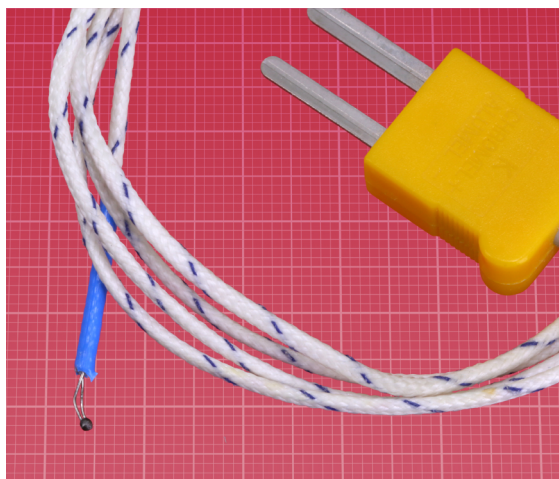


Figure 25-2 Overview of the thermocouple in the previous photograph.

A thermocouple sold as a commercial product is likely to be enclosed in a probe, as shown in [Figure 25-3](#).



Figure 25-3 A probe that contains a thermocouple.

Schematic Symbol

A schematic symbol that is often used to represent a thermocouple is shown in [Figure 25-4](#). Because this component does not consume current, the plus and minus signs do not mean that power should be applied to the wires. The positive sign indicates which wire will generate a higher voltage than the wire with the negative sign.

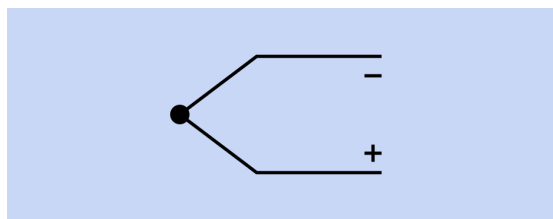


Figure 25-4 A symbol that is often used for a thermocouple.

Comparison of Temperature Sensors

In this Encyclopedia, temperature sensors are divided into five main categories, each of which has a separate entry. For convenience, a comparative summary is included in the entry for **NTC thermistors**. See [“Addendum: Comparison of Temperature Sensors”](#) for an overview. Also see [Figure 23-9](#).

Thermocouple Applications

Thermocouples have a wider range than any other form of contact temperature sensor, some types being capable of measuring up to 1,800 degrees Celsius. The main limitation is the ability of the joint between the wires to withstand the heat. Appropriate insulation must be used, but segments of ceramic tube are marketed to serve this purpose if necessary.

The very small thermal mass of a thermocouple enables a rapid response to temperature fluctuations. No self-heating occurs, because the thermocouple consumes no power. It is simple and robust. However, its response is very non-linear, and the tiny voltages involved are vulnerable to corruption by electrical noise. Accuracy is usually not better than plus-or-minus 0.5 degrees Celsius, and may be less at low temperatures.

Thermocouples are commonly found in laboratories and in some industrial applications, such as monitoring the temperature in a blast furnace or inside an internal combustion engine.

They may also measure temperatures as low as -200 degrees, but at temperatures below -100 degrees the temperature coefficient diminishes to the point where voltage increments are less than 30 μ V per degree Celsius.

How a Thermocouple Works

When one end of a piece of wire is maintained at a temperature that is different from the other end, the temperature gradient along the wire creates a small electromotive force that manifests itself as a difference in electrical potential between one end of the wire and the other. This is known as the *Seebeck effect*, named after the man who discovered it. The magnitude of the potential will depend on two factors: the temperature difference between the ends of the wire, and the type of wire that is used.

Figure 25-5 illustrates the concept. Part 1 of this figure shows two wires, named A and B. The left ends of the wires are heated to the same temperature, T_X , while the right ends remain at a cooler temperature, T_Y . Because the wires are composed of different metals, the voltage drop across each wire will be different.

To make this model useful, some factors must be eliminated. In part 2 of Figure 25-5, the hot ends of the wires have been welded together. This now guarantees that they share the same temperature and the same voltage, V_X . We do not yet know what these X values are.

In part 3 of the figure, the cold ends of the wires are clamped in an isothermal block, which keeps them at an equal temperature, still represented as T_Y . The block is not electrically conductive, so the cold ends of the wires still have different voltages, V_A and V_B . We cannot measure these voltages directly, because they are relative to V_X , which is unknown. However, a volt meter can measure V_A and V_B relative to each other.

The volt meter will have its own voltage gradient on its wires, and possibly a temperature

gradient too, but both of these wires are made of the same metal (probably copper) and share the same temperature gradient. Therefore, their effects will be equal.

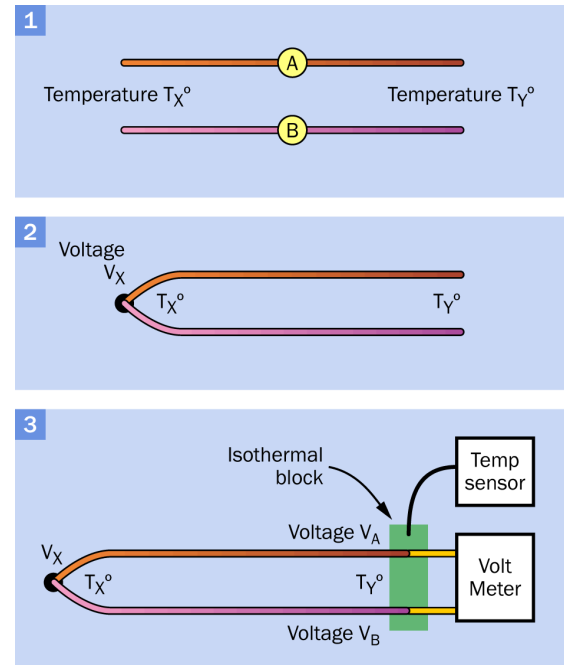


Figure 25-5 Basic principles of a thermocouple. See text for details.

A mathematical relationship exists between the temperature gradient and the voltage difference in each thermocouple wire. Suppose K_A is a constant or function that enables the voltage difference in wire A to be determined from its temperature gradient, and K_B serves the same function for wire B. Suppose T_{DIF} is the difference in temperature between T_X and T_Y . We may state:

$$K_A * (T_{DIF}) = V_X - V_A$$

$$K_B * (T_{DIF}) = V_X - V_B$$

By subtracting the second equation from the first and rearranging the terms, we get:

$$T_{DIF} * (K_A - K_B) = V_X - V_A - V_X + V_B$$

The two V_X terms cancel out, leaving $V_B - V_A$ on the right. We can recognize $V_B - V_A$ as the volt-

age difference measured by the meter. Call it V_M . So:

$$T_{\text{DIF}} = V_M / (K_A - K_B)$$

This enables calculation of the temperature difference between the ends of the wires, based on the meter reading and the conversion factor for each wire, which can be found experimentally. Because T_Y is being held at a known, constant value, we can determine the value of T_X :

$$T_X = T_Y + T_{\text{DIF}}$$

Thermocouple Details

When the thermocouple was first invented, the cold ends of the wires were placed in a bath of ice and water, forcing them to acquire and maintain a known temperature of 0 degrees Celsius.

The advent of accurately calibrated thermistors made it possible simply to measure the temperature of the cold ends. In this way, a thermistor enables a thermocouple to work. This prompts the question: why not just use the thermistor to measure T_X , and throw away the thermocouple? The reason is that a thermistor has a more limited range, seldom being used for temperatures above 150 degrees Celsius.

Note that the “hot end” of the thermocouple wires does not actually have to be hotter than the “cold end,” even though those terms are commonly used. The equation to find T_X works just as well if T_Y is higher than T_X . The temperature difference will simply have a negative value instead of a positive value.

Because “hot” and “cold” are misleading terms, modern documents generally refer to the “measurement junction” and the “reference junction” of the wires. Note, however, that the wires are not actually joined with each other at the reference junction.

A common misconception is that voltage is generated where the wires are joined at the measurement junction. This is not correct. The

voltage is a function of the temperature gradient between the measurement junction and the reference junction in each wire. Therefore, the way in which the wires are joined is irrelevant, provided there is an electrical connection between them. They can be welded, soldered, brazed, or crimped together.

How to Use It

Where a thermocouple is used in a laboratory, typically each wire is insulated, and they terminate in a plug that is inserted in a meter. The reference junction is hidden inside the meter, along with some electronics to decode the temperature data. The meter must have a setting that is appropriate to the specific type of thermocouple being used, so that the conversion factors are correct.

Because the type of metal in each wire must be consistent all the way from the measurement junction to the reference junction, other types of wires cannot be used to extend the reach of a thermocouple. Any extension must use wires made from the same metals. Connectors, also, must have pins and sockets that match the types of metals in the wires.

An extension wire for a thermocouple is shown in [Figure 25-6](#).



Figure 25-6 Extension wire for a type K thermocouple. Note the polarized plug.

Types of Thermocouples

Thermocouples are identified by ANSI standard codes consisting of single letters of the alphabet, shown below. Temperature ranges are approximate, in Celsius, with minimums rounded up and maximums rounded down to the nearest 50 degrees. Some data sources recommend narrower temperature ranges for practical use.

K type

-250 to +1,350 degrees. Most popular type of thermocouple. Positive wire is a nickel-chromium alloy, negative wire is a nickel-aluminum alloy. Commonly used in 3D printers.

J type

-200 to +1,200 degrees. Positive wire is iron, negative wire is a copper-nickel alloy. The iron wire is magnetic and vulnerable to corrosion. This thermocouple is not recommended for low temperatures, even though it is theoretically capable of measuring them.

T type

-250 to +400 degrees. Recommended for cryogenic applications. Positive wire is copper, negative wire is a copper-nickel alloy.

E type

-250 to +1,000 degrees. Most sensitive type, with the highest temperature coefficient. Positive wire is a nickel-chromium alloy, negative wire is a copper-nickel alloy.

N type

-250 to +1,300 degrees. An alternative to the K type, more stable at higher temperatures. Positive wire is a nickel-chromium-silicon alloy, negative wire is a nickel-silicon-magnesium alloy.

R type

-50 to +1,750 degrees. For high-temperature applications. Positive wire is

a platinum-rhodium alloy, negative wire is platinum. Very low temperature coefficient.

S type

-50 to +1,750 degrees. For high-temperature applications. Positive wire is a platinum-rhodium alloy, negative wire is platinum. Very low temperature coefficient.

Seebeck Coefficients

Datasheets for thermocouples list the Seebeck coefficient, which is the temperature coefficient caused by the Seebeck effect, measured in microvolts per degree. In other words, the value provided by a Seebeck coefficient is the number of additional microvolts that a thermocouple will generate for an increase in 1 degree Celsius.

Each type of thermocouple has a different coefficient, and because thermocouples tend to have a very nonlinear response, the coefficients will vary with temperature. [Figure 25-7](#) shows the variations for six types of thermocouples, over a range from -400 to +1,400 degrees Celsius. It is important to understand that the vertical scale shows the coefficient for each type of thermocouple—that is, the [change](#) in voltage, not the [actual](#) voltage, at temperature values along the horizontal axis.

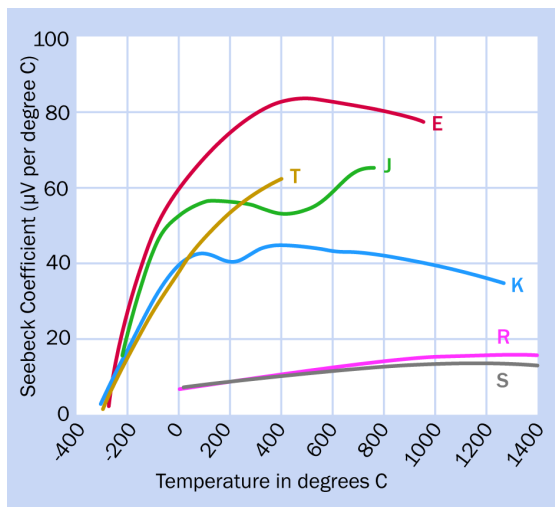


Figure 25-7 The Seebeck (temperature) coefficient for six types of thermocouples. Partially derived from a data-sheet published by Analog Devices.

R and S type thermocouples have a relatively consistent response, but cannot achieve great accuracy, because the voltage increment is so small for each change in temperature. The K type thermocouple does relatively well between 0 and 1,200 degrees, but the J type only performs adequately between 0 and 800 degrees, while the T and E types are quite inconsistent.

For lower voltage readings, electrical noise becomes an issue. Thermocouple wires are often twisted together, and may also be shielded, to reduce sensitivity to noise. The electronics that decode the thermocouple voltage should include a filter to reject 50Hz or 60Hz interference from wiring in the vicinity.

Chips for Output Conversion

Meters that are designed to interpret the output from a thermocouple and display a temperature tend to be expensive, and may not be convenient for a custom-built application. Fortunately integrated circuit chips are now available to amplify the thermocouple output and apply signal conditioning to create an almost linear response.

The AD8494 and AD8496 from Analog Devices are precalibrated by laser wafer trimming to match the characteristics of J type thermocouples, while the AD8495 and AD8497 match the characteristics of K type. The chips require a power supply of at least 3VDC and have an analog output of 5mV per degree Celsius, enabling measurement over a range of almost 1,000 degrees. They require a very low supply current of 180µA. The manufacturer claims an accuracy of plus-or-minus 2 degrees Celsius.

The chip contains a temperature sensor which should be at the same temperature as the reference junction of the thermocouple. This means that the reference junction (typically, at the socket where a thermocouple plug is inserted) should be as close to the chip as possible, and the chip should be protected from heat created by other components. Any difference in temperature between the reference junction and the chip will create a temperature measurement error.

The MAX31855K chip from Maxim Integrated Products is another thermocouple-to-digital converter. It linearizes the output from the thermocouple and digitizes it as a temperature that is accessible by a microcontroller via a serial SPI bus. Breakout boards with this chip are available. The last letter of the chip number specifies the thermocouple type. Variants for J, K, N, T, S, and R are available.

The AD8495 is mounted on a breakout board from Adafruit, and the MAX31855K is mounted on a breakout board from Sparkfun. These boards are pictured in [Figure 25-8](#).

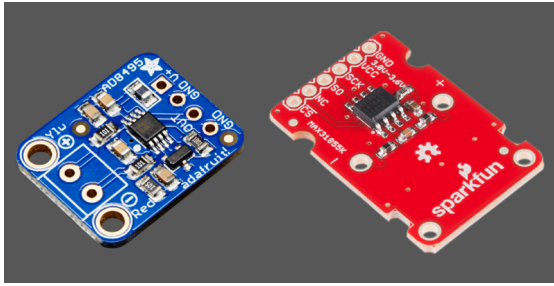


Figure 25-8 Thermocouple amplifier/conversion chips from Adafruit (left) and Sparkfun (right).

Thermopile

A thermopile is an assembly of thermocouples connected in series, as suggested in [Figure 25-9](#), where a hot area is shown on the left and a cooler area is shown on the right. The figure assumes that the orange-colored wires have a voltage difference of 5mV from left to right, as a result of the temperature difference, while the purple-colored wires only have a voltage difference of 1mV from left to right. Therefore, the voltage difference between each reference junction and the next is 4mV, as shown on the right side of the figure. So long as the temperature difference exists, the voltage differences will be cumulative, totaling 16mV between top and bottom in this example.

Note that the multiple thermocouple junctions are not electrically connected with each other in each temperature zone.

In reality, more thermocouples than this will be added, and the voltage differences may be lower.

Generally a thermopile is not sold as a separate component from retail suppliers, but is built into other devices. It may be used to generate small amounts of current from a heat difference, as in an infrared thermometer. It can also be used as a safety device to shut down a gas supply if a burner is not lit. See [Chapter 28](#).

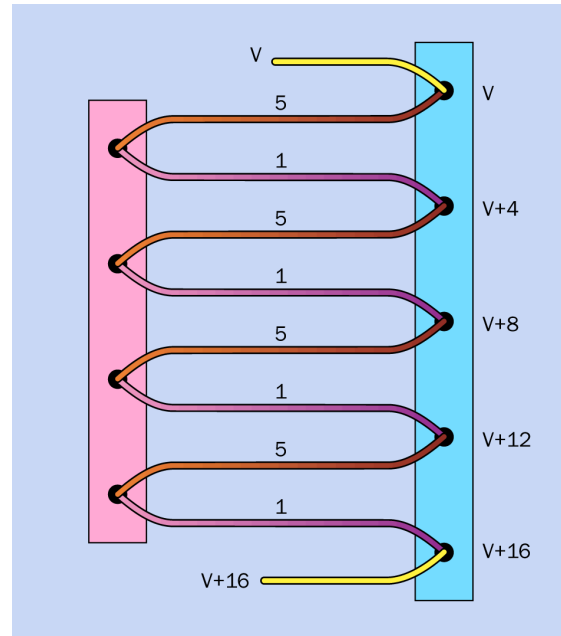


Figure 25-9 The operating principle of a thermopile. The numbers represent mV, but are chosen arbitrarily as an example.

What Can Go Wrong

Polarity

The output from a thermocouple has polarity. If this is not observed, an error will result.

Electrical Interference

Thermocouple wires are vulnerable to electrical interference, and should be a twisted pair or, ideally, shielded.

Metal Fatigue and Oxidation

The wire used in some thermocouples tends to be relatively brittle, and cannot withstand much flexing. Also, some metals or alloys are vulnerable to oxidation.

Using the Wrong Type

Different types of thermocouples have totally different characteristics. The electronics to decode the signal from a thermocouple must be matched to the type of thermocouple being used. The plugs on the ends of the thermocou-

ple wires are often retained with screws. A detached plug should be replaced immediately, to avoid the error of attaching it to the wrong type of thermocouple.

Heat Damage from Creating a Thermocouple

If a thermocouple is made from two wires on a DIY basis by welding the tips of the wires together, minimal heat must be used to avoid degrading the alloys in the wires.

RTD (resistance temperature detector)

RTD is an acronym, either for *resistance temperature detector* or *resistive temperature device*. No definitive information seems to exist regarding which term is correct, but *resistance temperature detector* is more common.

Occasionally an RTD may be described as a *PTC thermistor*, but its sensing element is different, consisting of pure metal wire or film.

OTHER RELATED COMPONENTS

- **thermocouple** (see [Chapter 25](#))
- **NTC thermistor** (see [Chapter 23](#))
- **PTC thermistor** (see [Chapter 24](#))
- **semiconductor** temperature sensor (see [Chapter 27](#))
- **infrared temperature** sensor (see [Chapter 28](#))

What It Does

A *resistance temperature detector*, also known as a *resistive temperature device*, is usually referred to as an **RTD**. It has a positive temperature coefficient (that is, its resistance increases as its temperature increases) but differs from a **PTC thermistor** in that its sensing element is pure metal instead of a semiconductor.

Comparison of Temperature Sensors

In this Encyclopedia, temperature sensors are divided into five main categories, each of which has a separate entry. For convenience, a comparative summary is included in the entry for **NTC thermistors**. See "[Addendum: Comparison of Temperature Sensors](#)" for an overview. Also see [Figure 23-9](#).

RTD Attributes

Positive attributes of RTDs include:

- Accuracy, often plus-or-minus 0.01 degrees Celsius. This very small tolerance allows excellent interchangeability.
- Stability, with a response that drifts by as little as 0.01 degree per year.
- The output is an almost linear function of temperature, making them easily used with a microcontroller.
- Immunity to electrical noise.
- Reasonably rapid response to temperature changes (about 1 to 10 seconds).

Undesirable attributes include:

- Temperature coefficient about one-tenth of that of an NTC thermistor.
- To measure the resistance, some current must pass through the sensor, raising the possibility of self-heating (as in the case of other temperature sensors, with the exception of a thermocouple).
- Relatively high cost, especially of the wire-wound type.

The resistance curves for three generic NTC thermistors are shown in Figure 26-1, plotted against the resistance for a generic platinum RTD that has a reference resistance of 100 ohms at 0 degrees Celsius. Note that this graph is unlike many that illustrate the response of NTC thermistors, in that its vertical scale is not logarithmic.

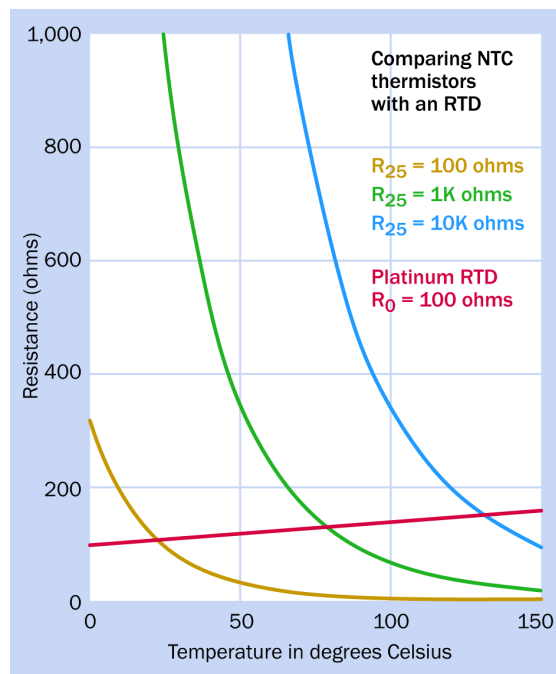


Figure 26-1 The brown, green, and blue curves show the resistance of three generic NTC thermistors varying with temperature. The red line shows the resistance of a platinum RTD. Derived from a diagram created by Texas Instruments.

Schematic Symbol

There is no specific schematic symbol for an RTD. Often the symbol for a thermistor may be used. See Figure 23-1.

Applications

Because of its high accuracy, an RTD may be used where precision is important. It can calibrate other temperature sensors, and may measure the temperature of a reference junction of a thermistor. However, it requires sensitive electronics for signal conditioning, because of its low temperature coefficient.

How It Works

An RTD exploits the fractional increase in electrical resistance of a metal film, metal filament, or (in some cases) a carbon film, when the temperature of the metal rises. In its simplest form, an RTD is a 2-wire device with no polarity.

The sensing element is often made from platinum, as this has a linear response to temperature over a wide range. High-quality RTD sensors with a wide temperature range usually consist of platinum wire that is wound around a glass or ceramic core. Smaller sensors may be fabricated from a thin layer of platinum evaporated onto an insulating substrate. Nickel may be substituted for platinum, and has a more sensitive but less linear response.

The wire-wound type is usable at temperatures as high as 500 Celsius (up to 1,000 degrees for some platinum-element types). Some variants are able to measure temperatures as low as -250 degrees.

DIN 60751 is an international standard defining the performance of platinum RTDs. It specifies a reference resistance of 100.00 degrees at 0 degrees, and a temperature coefficient between 0 degrees and 100 degrees. Outside of this range, a formula defines the response.

The response is almost precisely linear, ranging from 100 ohms at 0 degrees to approximately

138 ohms at 100 degrees. The temperature deviation from a straight line, between 0 and 100 degrees, is no more than plus-or-minus 0.8 degrees.

However, current through the RTD must be restricted to avoid self-heating. A range of 0.5mA to 1mA is recommended.

Variants

Some RTDs are potted in glass or resin, as shown in Figure 26-2. This shows an RTD in the TFPTL range from Vishay, containing a nickel thin-film sensing element with a temperature coefficient of about 0.4% and a tolerance of 0.01%. It is available with a very wide range of reference resistances, from 100 ohms to 5K (measured at 25 degrees Celsius). The temperature range is -55 to +70 or -55 to +150 degrees, and the maximum voltage is 30V to 40V, depending on the specific component.



Figure 26-2 An RTD in the TFPTL series from Vishay. The background grid is in millimeters.

The flat package shown in Figure 26-3 may be encased in a protective sheath of plastic or silicone rubber, and can be used for surface temperature sensing where the component is glued to the exterior of a flat-sided container. This figure shows an RTD in the L420 range

from Heraeus Sensor Technology, containing a platinum thin-film sensing element with a temperature coefficient of 0.385%. It is available in reference resistances of 100, 500, and 1,000 ohms (measured at 25 degrees Celsius). The temperature range is -50 degrees to +400 degrees.

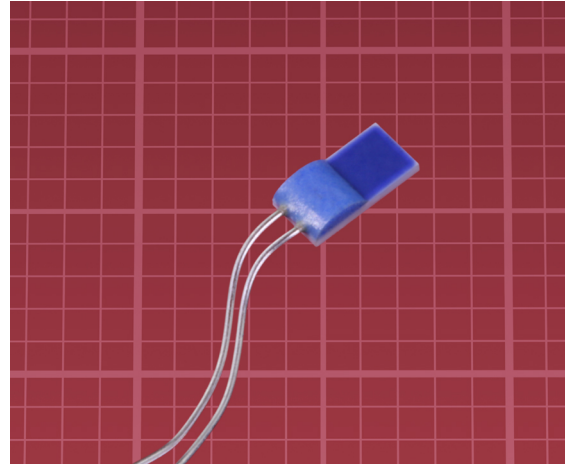


Figure 26-3 An RTD in the L420 series from Heraeus. The background grid is in millimeters.

Wiring

The leads to an RTD can be a source of error. If a simple 2-wire configuration is used, the leads will have an unknown resistance that will be affected by temperature, just as the element inside the RTD will have a temperature-sensitive resistance.

To enable temperature compensation, a three-wire design can be used. Figure 26-4 illustrates the principle. In the first section of this figure, resistances R_A and R_B remain unknown. In the second section, the resistances of R_B and R_C can be found by passing a test current through one wire and back through the other, bypassing the component. Assuming that all of the leads are identical in length and composition, the resistance of $R_A + R_B$ will be equal to that of $R_B + R_C$.

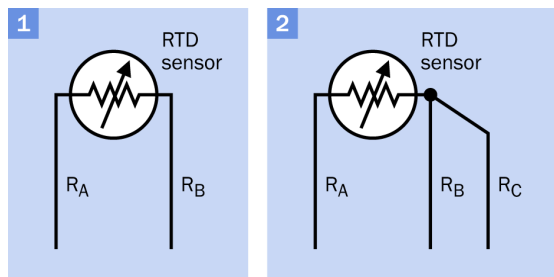


Figure 26-4 A three-wire configuration enables temperature compensation for the leads to an RTD. See text for details.

RTD Probe

For practical applications, an RTD sensor is often packaged inside a probe that can be indistinguishable from the type of probe used with a **thermocouple**. However, a thermocouple always uses two wires, as the wires themselves create the voltage. An RTD often uses three wires, as shown in Figure 26-5. This particular sensor is sold for use in a “Brew-Magic” system for brewing craft beer on a commercial basis.



Figure 26-5 A three-wire RTD is packaged inside this steel probe.

Signal Conditioning

To process the signal of an RTD, a chip such as the LM75 from National Semiconductor can be used. This is calibrated for connection with a platinum RTD. Internally it converts the resistance of an RTD to a value of 5mV per degree Celsius. This then passes through an analog-to-digital converter on the chip, creating a digital value that can be read via an I2C bus.

What Can Go Wrong

Self-Heating

Self-heating is an issue for RTDs, just as it is for thermistors. Current through an RTD should be limited to 1mA, especially when measuring low temperatures.

Insulation Affected by Heat

The resistance of insulation on the wires leading to a sensor can change with temperature, leading to incorrect resistance readings. This is a more likely source of problems for RTDs than for thermistors, as RTDs are often used at higher temperatures and have a lower temperature coefficient.

Incompatible Sensing Element

If signal conditioning is applied to an RTD that has an incompatible sensing element, temperature readings will be incorrect. For example, an RTD with a nickel element should not be used with signal conditioning designed for a platinum element.

semiconductor temperature sensor

27

This type of sensor may also be referred to as a *bandgap temperature sensor*, a *diode temperature sensor*, a *chip-based temperature sensor*, or an *IC temperature sensor*.

The unfortunate term *integrated silicon-based sensor* is sometimes used, which can be confused with *silicon temperature sensor* (also known as a *silistor*), which is a type of **PTC thermistor**. See Chapter 24.

Some vendors do not divide temperature sensors into clear categories. Semiconductor temperature sensors may be classified as *board-mount* temperature sensors, even though many of them have leads and are not specifically designed to be mounted on circuit boards.

A semiconductor temperature sensor with a digital output is sometimes described as a *digital temperature sensor* or *digital thermometer*. This can be misleading, as the outputs from other types of temperature sensors may be digitized with appropriate components.

OTHER RELATED COMPONENTS

- **thermocouple** (see Chapter 25)
- **NTC thermistor** (see Chapter 23)
- **PTC thermistor** (see Chapter 24)
- **infrared temperature** sensor (see Chapter 28)
- **RTD** (resistance temperature detector) (see Chapter 26)

What It Does

A **semiconductor** temperature sensor is an integrated circuit chip incorporating a sensing element composed of transistor junctions. It has an approximately linear response and is easy to use, in some cases being designed for direct connection with a microcontroller, requiring no additional components.

In analog variants, the output consists either of voltage or current that varies with temperature. These components have a positive temperature

coefficient, except for a few CMOS variants where a voltage output diminishes as temperature increases.

Digital variants are becoming more common, creating a numeric output accessible by a microcontroller.

In almost all semiconductor temperature sensors, the characteristics of silicon dioxide limit the temperature range to approximately -50 to +150 degrees Celsius (sometimes less).

This type of component is not (yet) as low in cost as a thermistor, but can include its own amplification, signal processing, and (optionally) analog-to-digital conversion on one chip.

Comparison of Temperature Sensors

In this Encyclopedia, temperature sensors are divided into five main categories, each of which has a separate entry. For convenience, a comparative summary is included in the entry for **NTC thermistors**. See “[Addendum: Comparison of Temperature Sensors](#)” for an overview. Also see [Figure 23-9](#).

Semiconductor Temperature Sensor Applications

When a semiconductor temperature sensor is used in surface-mount format, it can measure the temperature of the board on which it is mounted. This enables protection from overheating, often in power supplies.

Because the sensing elements and signal processing circuits are all chip-based, they can be transplanted into other types of sensors. For example, a gas pressure sensor or a proximity sensor can have onboard compensation using a semiconductor temperature sensor. They have also been built into computer CPUs such as the Pentium series from Intel.

Some variants are manufactured in a three-lead TO-92 package, appearing superficially similar to bipolar transistors. They are suitable for remote temperature sensing, and have automotive applications such as measuring the temperature of the transmission, engine oil, or cabin interior. They may also be found in some heating and air-conditioning systems, and some kitchen equipment.

Schematic Symbol

No unique schematic symbol has been developed for a semiconductor temperature sensor. It may be represented by a rectangle contain-

ing text abbreviations to represent pin functions, similar to other types of integrated circuit chips.

In the case of a sensor with an output consisting of current that varies with temperature, the sensor may be shown as a [current source](#), using the symbol in [Figure 27-1](#). However, this symbol is not specific to temperature sensors; it is used for any component that is a current source.

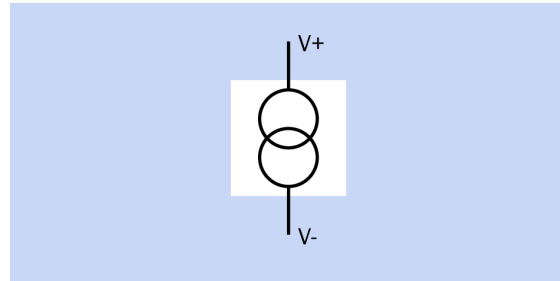


Figure 27-1 A temperature sensor whose output consists of current varying with temperature may be shown in a schematic as a current source, using this symbol.

Attributes

Desirable attributes of a semiconductor temperature sensor include:

- Easy to use. Few or no external components may be required, and little or no signal processing.
- Factory-calibrated, with an almost linear response.
- Versions with a digital output are easy to add to any system that already has an I2C bus. For additional details about protocols such as I2C, see [Appendix A](#).

Undesirable attributes of semiconductor temperature sensors include:

- Limited temperature range, the same as thermistors.
- Self-heating issues, especially in versions where signal-processing functions are built into the same chip.

- Not as rugged as some types of temperature sensors.

For an explanation of terminology used for temperature sensors in datasheets, see “[Thermistor Values](#)”.

How It Works

When a constant current is flowing through a p-n junction in a diode, the voltage across the diode will change by about 2mV for each change in temperature of 1 degree Celsius. This can be demonstrated by the simple circuit shown in section 1 of [Figure 27-2](#).

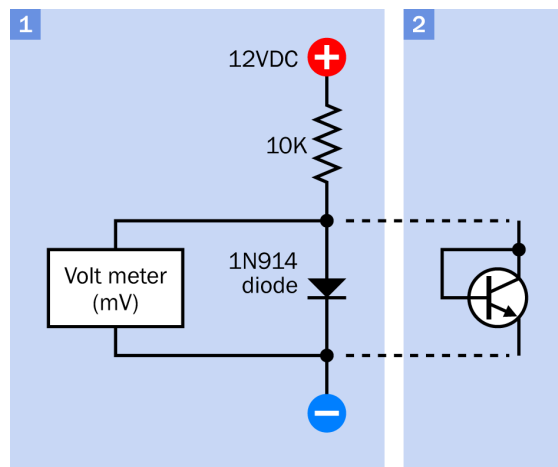


Figure 27-2 Left, a basic circuit for demonstrating the temperature sensitivity of a diode. Right, an NPN transistor can be substituted to emulate the diode.

Similarly, the voltage across the p-n junction in an NPN transistor varies with temperature, if the current is constant. A transistor can be substituted for a diode as suggested in section 2 of [Figure 27-2](#). Integrated circuit chips that contain transistors can measure temperature by exploiting this phenomenon.

CMOS Sensors

Some semiconductor temperature sensors use CMOS instead of bipolar transistors. The general concept is similar, but they are described

separately, below. See “[CMOS Semiconductor Temperature Sensors](#)”.

Multiple Transistors

The heat sensitivity of a bipolar transistor can be defined with an equation. If the base-emitter voltage is V_{BE} , q is the charge of an electron, k is a constant (known as Boltzmann’s constant), T is the temperature in degrees Kelvin (relative to absolute zero), I_C is the collector current, and I_S is the saturation current (which is less than I_C):

$$V_{BE} = ((k \cdot T) / q) * \log_e (I_C / I_S)$$

The term \log_e means, “the logarithm to base e of the expression in parentheses.”

Because k and q have known values, the base-emitter voltage turns out to be proportional to the logarithm of the collector current divided by the saturation current. However, the saturation current depends on the geometry of a transistor, and varies with temperature in a nonlinear way.

To eliminate the factor of saturation current, one transistor can be compared with another transistor that has a larger emitter area. This enables the derivation of a new equation that specifies temperature while getting rid of the troublesome saturation currents, with their nonlinear behavior.

However, it may not be easy to fabricate two transistors, in the same silicon chip, that have the same characteristics except that the emitter area of one is bigger than that of the other. It is much easier to add multiple transistors in parallel, each of them identical to the first. The total emitter area will then be equal to the area in one transistor multiplied by the number of transistors.

In [Figure 27-3](#), assuming all the transistors are at the same temperature in addition to being of identical specification, we can now write two equations. The figure allows room to show only three transistors, but suppose there are N of

them. If V_{BE0} is the base-emitter voltage of transistor Q0, on the left, and V_{BEN} is the aggregate base-emitter voltage of the N transistors on the right:

$$V_{BE0} = ((k*T) / q) * \log_e (I_C / I_S)$$

$$V_{BEN} = ((k*T) / q) * \log_e (I_C / N*I_S)$$

From these, an equation can be derived that gets rid of I_C and I_S :

$$V_{BE0} - V_{BEN} = ((k*T) / q) * \log_e (N)$$

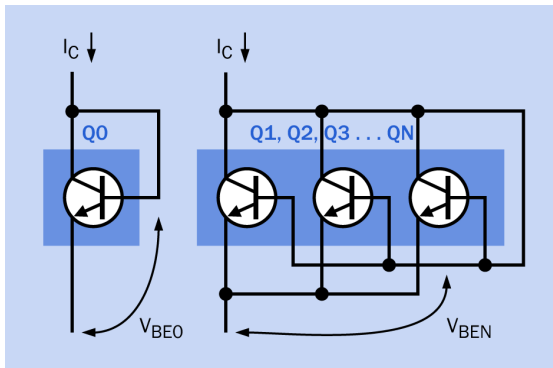


Figure 27-3 Comparing the base-emitter voltage of one transistor with a set of identical transistors can enable measurement of temperature regardless of the collector current and saturation currents, so long as all the transistors are at the same temperature. See text for details.

PTAT and the Brokaw Cell

Now if a comparator is added to control the current, a circuit known as the Brokaw Cell is created, shown in [Figure 27-4](#). This is also known generically as a [bandgap temperature sensor](#). (A couple of resistors have been omitted for the sake of simplicity.)

Typically, $N = 8$. That is, there is a set of eight transistors in addition to Q0 (only three being shown here). The voltage difference in the previous equation, $V_{BE0} - V_{BEN}$, now appears across R2 in the figure, and the voltage across R1 is [proportional to absolute temperature](#), often referred to by its acronym, [PTAT](#). This voltage can be found from this equation:

$$V_{PTAT} = ((k*T)/q) * \log_e(N) * (2*R1/R2)$$

The Brokaw Cell was the basis of the AD580 chip introduced in 1974 by Analog Devices, and the principle is now used very widely in semiconductor temperature sensors.

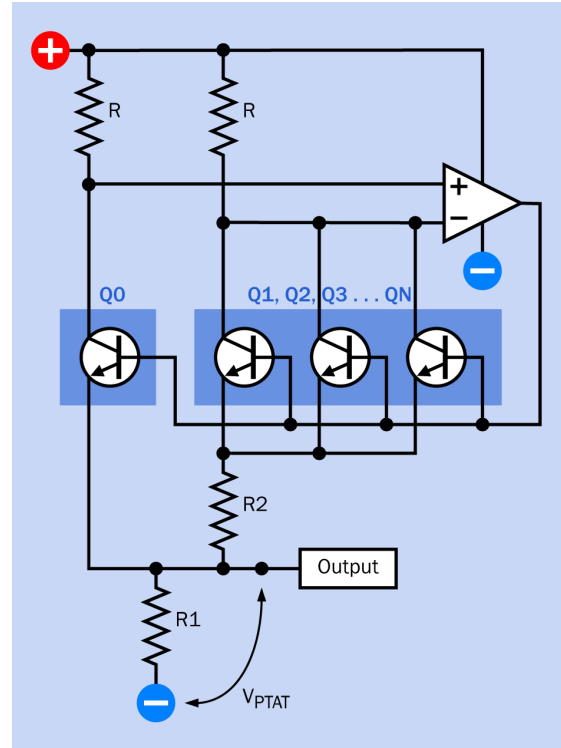


Figure 27-4 The Brokaw Cell. See text for details.

Variants

Three output types are used:

- Analog voltage output (voltage varies with temperature).
- Analog current output (current varies with temperature).
- Digital output.

A fourth type creates an output in the form of a square wave, either with a frequency or wavelength proportional to temperature. The Maxim MAX6576 and MAX6577 are examples. How-

ever, this type of output is so rare, it is not described in detail here.

Some semiconductor temperature sensors are CMOS-based, and have a voltage output with a negative temperature coefficient. They are described separately. See “[CMOS Semiconductor Temperature Sensors](#)”.

Analog Voltage Output

LM35 Series

The LM35 is a typical, widely used semiconductor temperature sensor, available from Analog Devices, Texas Instruments, and other manufacturers. Its output voltage changes by 10mV per degree Celsius over a range of approximately -50 degrees to +150 degrees. Accuracy is stated to be plus-or-minus 0.25 degrees at room temperature and plus-or-minus 0.75 degrees over the whole range.

The sensor can be obtained packaged like a transistor, in a TO-92 plastic capsule or metal can. It is also available as a surface-mount component, or in a TO-220 package, like a 5V voltage regulator, as shown in [Figure 27-5](#).

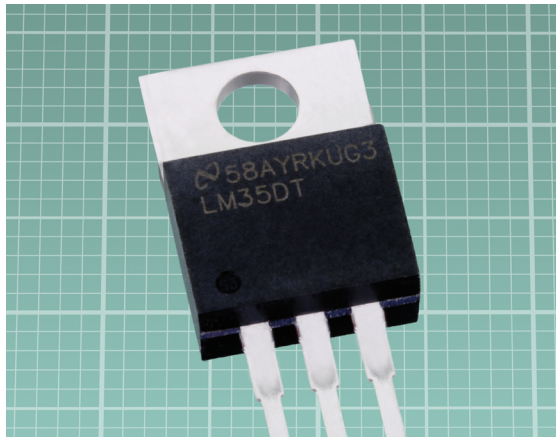


Figure 27-5 This version of the LM35 can be used to measure surface temperature when secured with a bolt. The background grid is in millimeters.

This is a three-wire component, two pins or solder pads being used for the power supply while the third serves as the sensor output. The sup-

ply voltage typically ranges from 4V to 30V. Necessary current consumption is only 60μA, which minimizes self-heating.

Because this device is specifically designed for the Celsius temperature range, its output is scaled to 0mV at 0 degrees. A pulldown resistor can be added to measure temperatures below zero.

A bypass resistor of 200 ohms between the output and ground is recommended as a precaution against capacitive effects in the cable run.

The LM34 is almost identical to the LM35, except that its output changes by 10mV per degree Fahrenheit instead of 10mV per degree Celsius.

LM135 Series

Although this sensor contains multiple NPN junctions, the manufacturer describes it as behaving like a zener diode in which the breakdown voltage is directly proportional to absolute temperature. The output increases by 10mV per degree over a range from -55 to +150 degrees Celsius.

For the LM135, the manufacturer claims an error of less than plus-or-minus 1 degree Celsius between 0 and 100 degrees. For the LM235 and LM335, in the same product series, the temperature range is narrower, the accuracy is lower, and the price, also, is lower. An LM335 sensor is shown in [Figure 27-6](#).

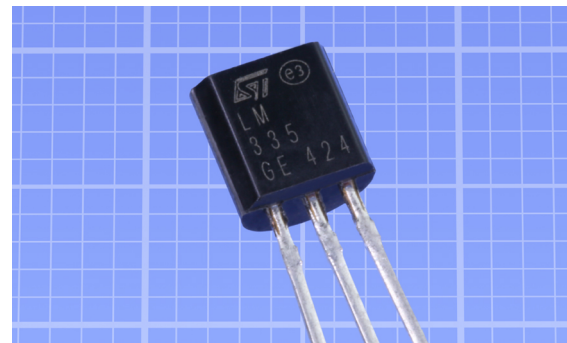


Figure 27-6 A sample of the LM335 temperature sensor in a TO-92 package. The background grid is in millimeters.

The sensor is available in a TO-92 package (plastic, like a transistor) or a TO-46 (metal can). It is also manufactured in a surface-mount format. The negative terminal is connected directly to ground, while the positive terminal is connected through a series resistor to the positive side of a power supply that can range from 5V to 40V. The third terminal, labeled “ADJ” on datasheets, allows for the output to be adjusted. [Figure 27-7](#) shows the basic circuit. The value of R1 can be chosen to establish an optimal current of 1mA through the sensor, although a range of 400 μ A to 5mA is tolerable.

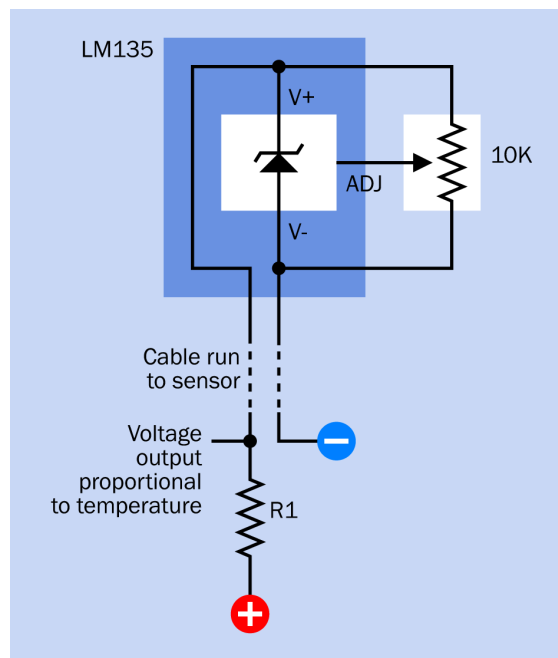


Figure 27-7 Basic schematic for using an LM135, including output adjustment. Because the sensing element behaves like a zener diode, it is represented with the zener symbol.

Analog Current Output

Fewer components exist using output current to measure temperature. The output is applied to a grounded resistor, and the voltage across the resistor then changes with current from the sensor.

The useful aspect of a current output is that its accuracy is unaffected by a cable run as long as 200 or 300 feet. Therefore, this type of component is appropriate as a remote sensor.

LM234-3 Series

This is a three-wire sensor, two wires being used for bias power supply and ground (labeled V+ and V- on the datasheet) and a third (labeled R) that delivers current proportional to temperature. The current from pin R passes through an external resistor to ground, and the voltage across this resistor varies by 214 μ V per degree Kelvin. A bias voltage ranging from 1V to 40V is required.

A sample of the LM234 is shown in [Figure 27-8](#).

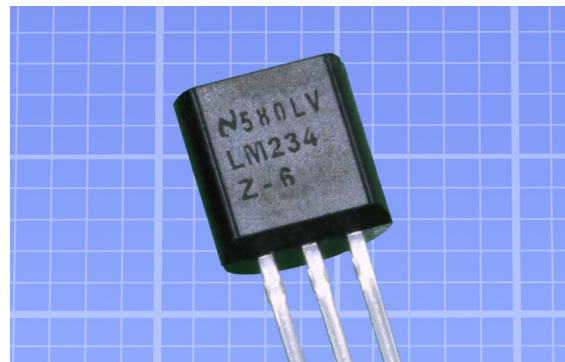


Figure 27-8 The LM234Z temperature sensor in a TO-92 package. The background grid is in millimeters.

If the component is used for remote sensing, the resistor should be 230 ohms and can be connected directly between pin R and pin V- of the sensor at the far end of a wire run. At the “home” end of the wire run, temperature output is taken from above a 10K resistor that is placed between the return wire and ground, as shown in [Figure 27-9](#). With these component values, the output voltage will change by 10mV per degree Kelvin.

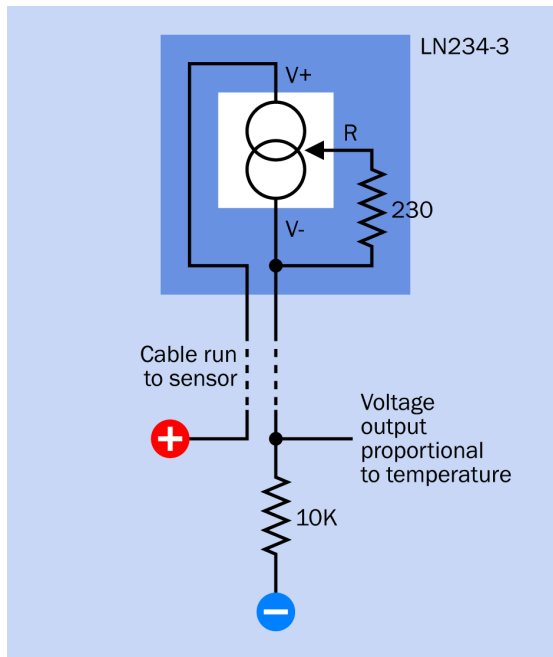


Figure 27-9 Using an LM234-3 sensor with a current output that varies with temperature.

The LM234-3 can be encapsulated in a plastic TO-92 package or a TO-46 metal can. A surface-mount version is also available.

The claimed accuracy is plus-or-minus 3 degrees. The temperature range is -25 degrees to +100 degrees.

AD590 Series

The Analog Devices AD590 (successor to the original AD580) is a current-output sensor that uses only two wires. Like the LN234-3 it is available in a TO-46 metal can, but with one lead making no internal connection. It also can be bought in a two-wire "flatpack," or as a surface-mount chip (with eight solder pads, only two of which are connected).

Using a supply voltage of 4V to 30V, the sensor's high-impedance output changes by 1 μ A per degree Kelvin. Voltage supply variations produce very small errors in the output current; substituting 10V for 5V creates a deviation of only 1 μ A.

Figure 27-10 shows an application for the AD590 using a resistor and a trimmer to adjust the scale factor. When properly set up, this circuit provides an output that changes by 1mV per degree Kelvin.

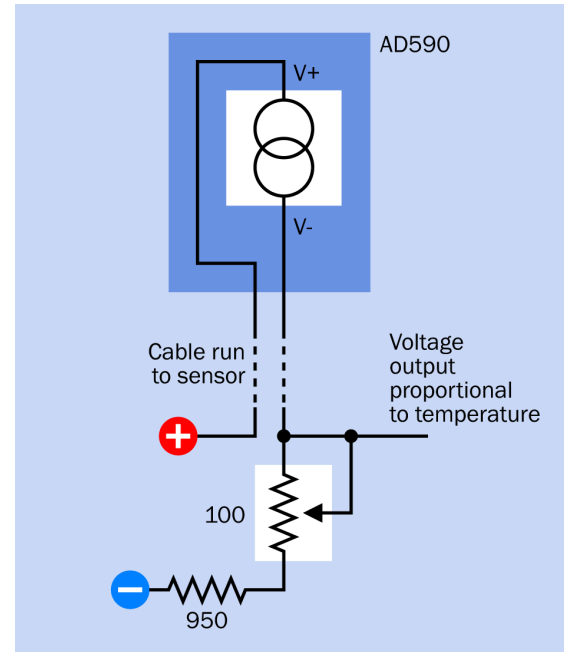


Figure 27-10 The simplest circuit allowing fine adjustment for use with an AD590 sensor.

Digital Output

Some of the most popular examples of semiconductor temperature sensors with digital output include the TMP102 series from Texas Instruments, MCP9808 series from Microchip, LM73 series from Texas Instruments and National Semiconductor, and DS18B20 series from Maxim. All of these components can measure a range typical of semiconductor temperature sensors, from approximately -50 to +150 degrees Celsius. Most of them claim an accuracy in the region of plus-or-minus 1 degree over the full range or 0.5 degree in the 0 to 100 range. With the exception of the Maxim DS18B20, which uses its own unique protocol, the components communicate via either the I2C or SMBus protocol.

TMP102 Series

This is available only in a surface-mount format. It has fewer features than the other sensors listed here, and is less accurate, claiming plus-or-minus 3 degrees Celsius over its maximum range of -40 to +125 degrees Celsius. However, it is less expensive. (For greater accuracy, the TMP112 is available.) This is a low-voltage chip requiring 1.4V to 3.6V as its power supply, drawing a quiescent current of 10 μ A. Temperatures are stored in a 12-bit or 13-bit format that requires some conversion, as a single bit represents 0.0625 degrees Celsius. An alert pin is activated when the measured temperature deviates above or below limits that are preset by the user. No hysteresis adjustment is available for the alerts. The TMP102 is available on a breakout board from Sparkfun, as shown in Figure 27-11.



Figure 27-11 The Texas Instruments TMP102 on a breakout board from Sparkfun.

MCP9808 Series

This multifunction sensor is available either as a regular surface-mount component, or as a surface-mount with an exposed “thermal pad.” It conforms with the I2C bus standard at up to 400kHz, allowing up to 16 sensors to share the same bus. The chip has a variety of temperature alert features, including high and low limits that can activate a dedicated “alert” pin, and a hysteresis value that can be set for the limits, to ignore brief temperature excursions. The chip can be put into “comparator mode,” where it simply provides logic-high or logic-low output

if the temperature is above or below a user-specified value. This feature makes the chip operate as a thermostat. Temperature resolution is user-selectable. The temperature storage format requires some conversion to obtain a Celsius value, to deal with negative values and fractional values. However the chip is available on a breakout board from Adafruit, as shown in Figure 27-12, and an Arduino code library is available.



Figure 27-12 The Microchip MCP9808 on a breakout board from Adafruit.

LM73 Series

This sensor is only available in surface-mount format. It conforms with the I2C bus standard at up to 400kHz. Its temperature resolution can be set to 11, 12, 13, or 14 bits. An “alert” pin becomes active if the temperature exceeds a preprogrammed limit. An “address” pin can select one of three device addresses by being held in logic-high, ground, or disconnected status. The chip can be put into shutdown mode when power conservation is necessary.

DS18B20 Series

Unlike most digital sensors, this is a three-wire component, because it uses Maxim’s proprietary “1-wire bus” with a unique protocol. The bus allows access to a 2-byte register storing digital output from the temperature sensor, but also allows the user to perform other functions, such as setting the resolution of an onboard analog-to-digital converter (which has a maximum resolution of 12 bits), setting a high-temperature and low-temperature alarm, and

allowing the sensor to be identified, as each component has a unique 48-bit serial number in ROM.

The chip can draw sufficient power from the data bus to operate, so long as the bus is held high by a 4.7K pullup resistor. (Maxim describes this as “parasite power.”) An internal capacitor sustains the chip briefly while the bus is used for its normal purpose of transferring data, but if the bus has low voltage for more than 480μs the chip will reset itself. The “parasite power” feature also will not work above 100 degrees Celsius. Perhaps recognizing that this system may create more problems than it solves, Maxim has also given the component a normal power input pin.

The DS18B20 is available in the TO-92 package and two sizes of surface-mount chip. Its lack of a standard I2C bus, and its use of complicated proprietary codes, create a steep learning curve. Still, it remains a popular sensor, and an Arduino code library for it is available online.

CMOS Semiconductor Temperature Sensors

CMOS variants of semiconductor temperature sensors have appeared relatively recently compared with the bipolar variants. They draw a very low quiescent current (typically, a few microamps) and can work with a power supply from 5.5VDC down to 2.2VDC, making them suitable for handheld battery-powered devices. An analog output is common. Popular examples are the LM20 and the LMT86 series.

Like bipolar sensors, the LMT86 sensors have a limited temperature range, between approximately -50 and +150 degrees Celsius. Again, like the bipolar sensors, they are available optionally in TO-92 and surface-mount packages. A significant difference is that the output has a negative temperature coefficient, diminishing by 10mV per degree Kelvin, because of the characteristics of CMOS semiconductors.

The claimed accuracy is plus-or-minus 0.25 degrees Celsius. The output voltage covers a range of about 2V, diminishing from 0.5V below the supply voltage at -50 degrees Celsius.

A sample of the LMT86 is shown in Figure 27-13.

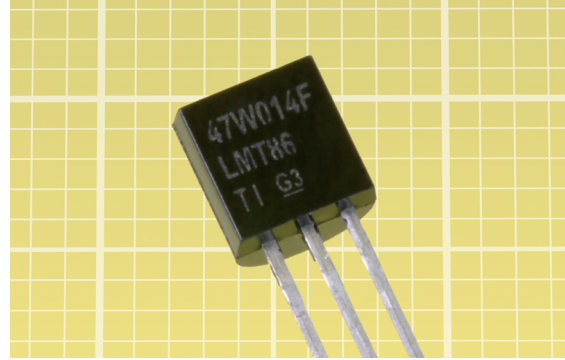


Figure 27-13 A sample of the LMT86 CMOS temperature sensor. The background grid is in millimeters.

What Can Go Wrong

Different Temperature Scales

Some voltage-output sensors create an output convertible to degrees Kelvin, while others use degrees Celsius. While the temperature degrees are the same in each scale, a component with an output in millivolts may assign 0mV either to 0 degrees Celsius or 0 degrees Kelvin (equivalent to -273.15 degrees Celsius). The advantage of using a Kelvin scale is that it avoids the problem of negative temperature values.

Rarely, a sensor may use degrees Fahrenheit.

Interference in Cable Runs

Sensors with a voltage output are susceptible to electrical interference. Twisted-pair or shielded cable runs are recommended when sensors are placed remotely.

For the Maxim DS18B20, which uses a 1-wire bus, multiple sensors should be connected along one run of wire (linear topology) instead of each sensor being connected to a central point (star topology). If the cable lengths are

longer than a few meters, the topology starts to matter.

Latency

The packaging of semiconductor temperature sensors can create latency in their response time. While a thermocouple consists only of a pair of wires joined by a small dot of melted metal, a TO-92 semiconductor package adds thermal mass that will slow the response considerably. Moreover, copper leads will conduct heat from a circuit board if the board is warmer than its environment.

Surface-mount chips have a very low mass, but must be soldered to some kind of board, even if it is a very small one.

Generally, other types of sensors may be appropriate where rapid response is necessary.

Processing Time

In a sensor with a digital output, the onboard analog-to-digital converter will add a small delay before the data becomes available, and during that delay, the component cannot respond to a new temperature. The output from an analog device may be more suitable for rapid detection of temperature variations.

infrared temperature sensor

An **infrared temperature** sensor is sometimes described as a *thermopile*. In reality the sensor module *contains* a thermopile. In this Encyclopedia, a thermopile is considered to be a separate component, described in the entry discussing **thermocouples**. See “Thermopile” in Chapter 25.

Other terms that are sometimes used for an infrared temperature sensor are *contactless thermometer* or *infrared thermometer*. This Encyclopedia classifies a thermometer as a commercially marketed product, not a component.

Devices such as a *radiation pyrometer*, *IR pyrometer*, *optical pyrometer*, or *thermal imager* provide ways of measuring infrared radiation, but are outside the scope of this Encyclopedia.

A **passive infrared** motion sensor (PIR) can detect infrared radiation, but only responds to fluctuations in intensity. An infrared temperature sensor measures the steady-state value of incident radiation.

OTHER RELATED COMPONENTS

- **passive infrared** motion sensor (see [Chapter 4](#))
- **thermocouple** (see [Chapter 25](#))

What It Does

Most temperature sensors discussed in other entries in this Encyclopedia are *contact sensors*, meaning that to measure the temperature of an object, liquid, or gas, they must make contact with it. In situations where contact is not possible or desirable, an **infrared temperature** sensor can be used. It responds to the *black-body radiation* (sometimes known as *characteristic radiation*) that is emitted by all materials above absolute zero (0 degrees Kelvin). This varies with temperature as a result of the movements of molecules.

Situations where noncontact sensors may be preferable to contact sensors include:

- An object is inconveniently located or too far away.
- The temperature of a large area must be measured.
- Contact with a small object would change the temperature of the object. The act of measurement would change the value being measured.
- The object is corrosive, abrasive, or otherwise liable to damage a sensor.
- The object is moving or vibrating.
- The surface of the object must not be contaminated (for example, unprotected foods).

- The temperature of the object is lower than around -50 degrees Celsius or higher than 1,300 degrees Celsius.

However, noncontact sensors have some limitations:

- Normally, only the surface temperature of a target can be measured.
- The optics of the sensor must be protected from dust, dirt, and liquid.
- The target must be clearly visible, in line-of-sight.
- Air pollution will degrade the temperature measurement. Some gases, such as carbon dioxide, will tend to absorb infrared radiation.
- The sensor will be affected by other sources, including reflected, transmitted, and convective heat.
- While an infrared sensor can theoretically respond to a very wide range of temperatures, in practice separate sensors of differing sensitivity are needed to cover a full range.
- Different types of materials emit differing intensities of black-body radiation, even if they are at the same temperature. Some compensation is necessary, or the surface of the object may have to be painted.

Applications

Handheld contactless thermometers were an early application for noncontact sensors.

In astronomy, the thermal radiation from the Sun and other stars is of interest to astronomers.

More recently, the declining cost of an infrared temperature sensor, and the ease of deploying it, have made it appropriate in consumer products. A significant area of adoption is in notebook computers and handheld devices, where

processor performance must be balanced against the need to prevent the case from becoming too hot to hold comfortably. In this kind of application, gluing a sensor to the interior of the case would be a manual operation during the production process and would require a wired connection. An infrared temperature sensor mounted on the circuit board, viewing the underside of the case, can achieve the same objective more simply.

A contactless sensor is also very useful for measuring the temperature of rotating objects, such as heating rollers in a laser printer.

Schematic Symbol

No specific schematic symbol exists for an infrared temperature sensor.

How It Works

While nanometers (abbreviated nm) are generally used to measure visible wavelengths, the longer wavelengths of far-infrared are often measured in micrometers (abbreviated μm). The measurable infrared values are defined as ranging from 0.7 μm to 14 μm , corresponding to peak emissions from a black body ranging in temperature from 200 degrees Kelvin to 6,000 degrees Kelvin (about -70 to +5,700 Celsius).

Unfortunately an object does not emit just one wavelength of black-body radiation for each temperature value. It emits a spread of wavelengths that becomes wider as the temperature increases. However, the peak intensity also increases with temperature, when measured as *spectral radiance*, which is defined as watts per steradian, per micrometer of wavelength. (A steradian is the solid angle at the top of a cone, in this case the cone being of emitted radiation.) Because the intensity increases, it can be used to calculate the temperature.

Figure 28-1 illustrates this concept. Note that both of the axes have logarithmic scales.

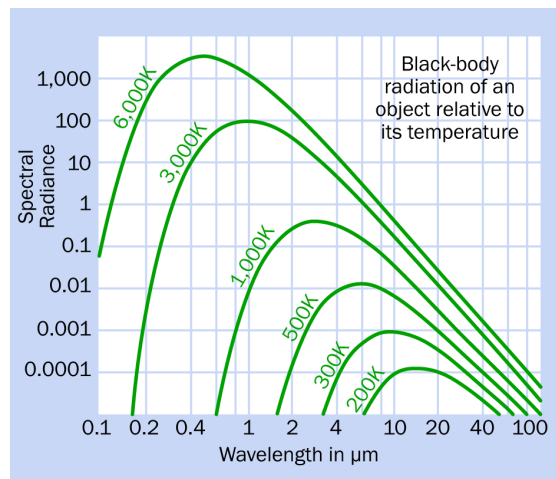


Figure 28-1 The increase in intensity, and the widening spread of wavelengths, of black-body radiation emitted by one object at six different temperatures (in degrees Kelvin).

All the curves are for one object. Each curve is specific to one temperature in degrees Kelvin, showing how the intensity of emitted radiation varies with wavelength. Note that radiation at wavelengths shorter than $0.7\mu\text{m}$ is within the visible spectrum; thus, objects at 1,000 degrees Kelvin, or hotter, may be seen to glow visibly.

Because of the wide variety of intensities and temperatures, an infrared sensor that is ideal for measuring a temperature of 1,000 degrees Kelvin will not provide an accurate result at 200 degrees Kelvin. The peak spectral radiance at 1,000 degrees is more than 10,000 times the peak at 200 degrees. Also, the curves in the figure are for an “ideal” object emitting pure black-body radiation. In reality, glass, plastic, and many other materials have a much lower [emissivity](#), meaning that they emit less radiation, and are classified as [gray bodies](#). A metallic object with a polished surface may emit one-tenth of true black-body radiation.

These issues cannot be ignored, but may be dealt with by relatively simple strategies. Infrared sensors will be rated according to their suitability for different temperature ranges, and the emissivity of the object being measured can be

determined by consulting standard tables. Alternatively, the object can be spray-painted with special black paint (such as “Senotherm” or “3-M Black”) that has a known emissivity of about 0.95 of pure black-body radiation. Alternatively, a specially formulated black sticker may be applied to the object that is being measured, so long as its temperature is within reasonable limits.

However, a basic infrared temperature sensor will not function reliably if it is pointed randomly at a variety of objects that vary widely in temperature. Specialized, expensive industrial devices incorporate compensation to deal with these issues, but they are outside the scope of this Encyclopedia.

Thermopile

A typical low-cost, chip-based infrared temperature sensor contains a [thermopile](#) consisting of multiple thermocouples etched into silicon and connected in series. The concept of a thermopile is illustrated in [Figure 25-9](#), where a brief explanation is included.

The configuration of the thermopile is arranged so that the hot junctions of its thermocouples are all clustered in a small central area, where they receive incoming radiation through a window (often made of silicon) that is transparent to infrared wavelengths. The cold junctions are dispersed around the periphery, where they are shielded from incoming radiation. One way to visualize this is shown in [Figure 28-2](#), although this is not a literal depiction of an actual sensor.

Instead of using alternating types of wire, as in a **thermocouple**, a chip-based thermopile often uses alternating segments of n-type and p-type silicon. The hot junctions are mounted on a thin film that has very little heat capacity, while the cold junctions are mounted on a thicker substrate that acts as a heat sink.

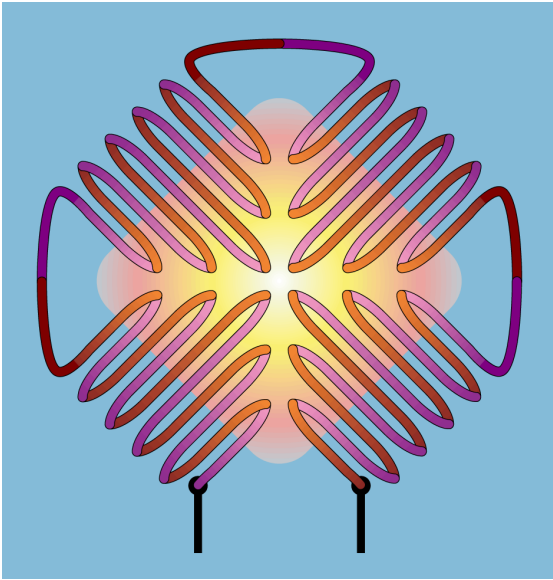


Figure 28-2 Simplified diagram of the thermopile configuration inside an infrared sensor chip. Radiation arriving through a window in the chip affects the thermocouple junctions in the central region, while junctions around the edges remain at a lower temperature.

Temperature Measurement

The voltage generated by the thermopile is related to the difference in temperature between the hot and cold thermocouple junctions. Thus there are three interrelated variables: hot temperature, cold temperature, and voltage. To calculate one variable, we must know the other two.

The hot temperature is what we wish to know. Therefore we must establish the voltage (which can easily be measured) and the cold temperature. The cold temperature can be determined by adding a thermistor inside the chip.

Typically an infrared temperature sensor with an analog output will have two pins that provide access to the internal thermistor, so that its temperature can be calculated from its resistance. Another two pins provide the voltage between the ends of the thermopile.

Interpreting and reconciling these values is not a trivial matter, especially bearing in mind that

the thermistor has a negative temperature coefficient and a nonlinear output, and the thermopile will also have some nonlinearity. To simplify this situation, some infrared temperature sensors incorporate electronics to perform the necessary calculations and provide a digital output. This output can be converted to degrees of temperature by some fairly simple mathematical operations in an external microprocessor.

Variants

Two types of sensors are popular. One is surface-mounted, such as the TMP006, shown in [Figure 28-3](#). This type generally has a digital output. The other type is a discrete component with four leads, such as the Amphenol ZTP135, shown in [Figure 28-4](#). Discrete components may have either an analog or a digital output.

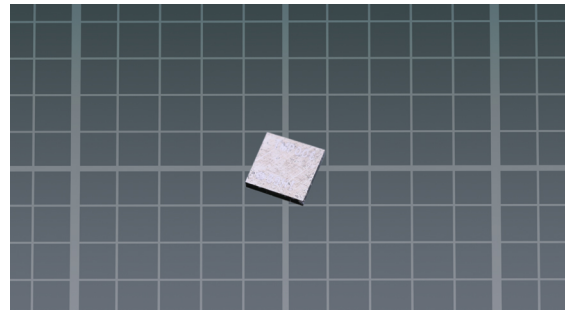


Figure 28-3 A surface-mount infrared temperature sensor with digital output. Eight tiny solder pads are located on the underside. The background grid is in millimeters.

Both types of sensors allow infrared light to enter through an area that is opaque to the visible spectrum but transparent to the appropriate range of wavelengths.

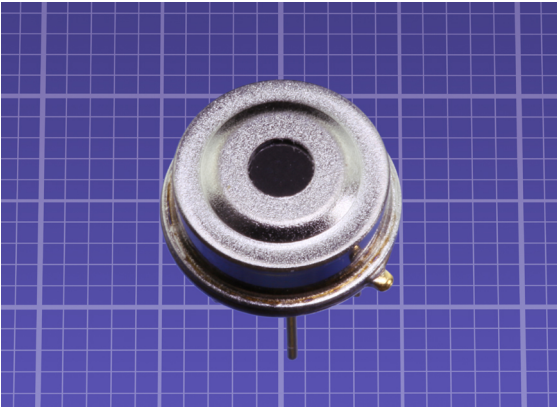


Figure 28-4 A through-hole infrared temperature sensor with analog output. The background grid is in millimeters.

The ZTP135 has an analog output shown in [Figure 28-5](#).

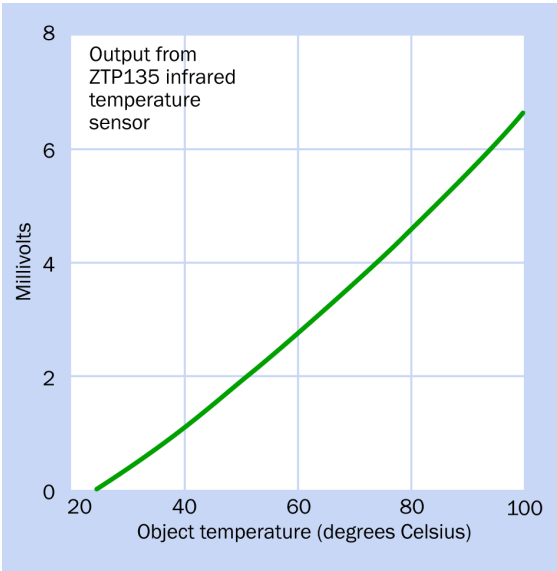


Figure 28-5 Analog output from an infrared temperature sensor.

The TMP006 is only about 1.5mm square, but is available on a breakout board from Sparkfun. Its successor, the TMP007, is available on a breakout board from Adafruit, shown in [Figure 28-6](#).

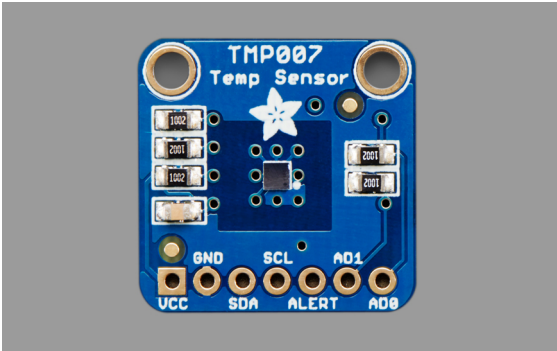


Figure 28-6 The TMP007 sensor mounted on a breakout board from Adafruit.

Surface-Mount Specifications

The TMP006 and TMP007 require a supply voltage that can range from 3.3V to 5V. These chips support the SMBus and I2C bus protocols, using a bus address that can be user-selected. An internal analog-to-digital converter uses 1 least-significant bit to represent 1/32nd of a degree Celsius, and data is saved as a 14-bit signed integer. Up to 16 temperature samples can be averaged internally.

Measurable temperature range is -40 to +125 degrees Celsius. Some hysteresis is built in. The TMP007 supports an alert mode if temperature falls above or below a user-specified threshold.

Sensor Arrays

Using multiple thermopile sensors arrayed in a line or a grid, with an array of lenses, it is possible to capture an image of temperature variations over a surface or a scene. This is known as [thermal imaging](#). It can detect heat leakage from buildings, indicating poor insulation, or can locate hot spots in electronic circuits. Hei-mann Sensor has pioneered the miniaturization of a 31 x 31 grid of thermopile sensors in a single TO-8 or TO-39 package.

Values

Temperature Range

Chip-based infrared temperature sensors are typically designed for a range from about -20

degrees Celsius to about +125 degrees Celsius. Their peak sensitivity is to wavelengths between 4μ and 16μ .

Other types of infrared temperature sensors can have a much wider temperature range, but are more costly.

Field of View

Often referred to by its acronym, *FOV*, the field of view is the angle at the apex of an imaginary cone extending outward from the sensor, defining a boundary where the sensitivity diminishes below 50% of the value directly in front of the sensor. Greek letter ϕ may be used to represent the angle between the surface of the cone and the center line, while θ represents the angle between the opposite surfaces of the cone (i.e., $2 * \phi$). This is shown in Figure 28-7. θ is usually the angle defined as field of view.

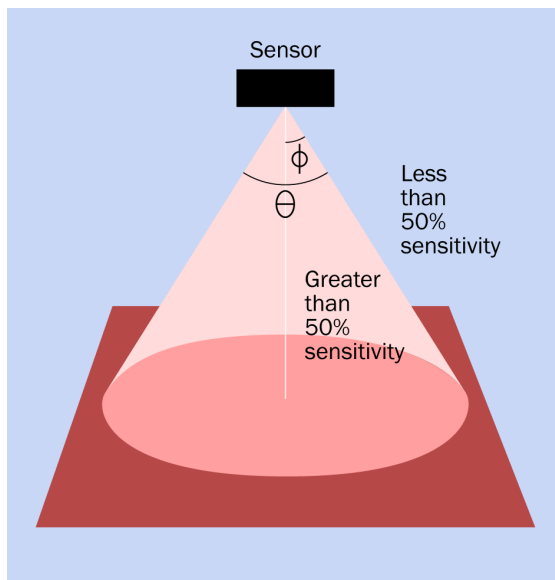


Figure 28-7 Measurement of the field of view from a sensor, defined as the boundary of an imaginary cone where sensitivity drops below 50%.

Because most infrared temperature sensing components do not have a lens, they are sensitive over a wide angle. The field of view is typically 90 degrees.

What Can Go Wrong

Inappropriate Field of View

The object being evaluated must fill the field of view of the sensor, to avoid measuring other objects around it.

Reflective Objects

A reflective object has lower infrared emissivity, and also may provide misleading output if the sensor is actually measuring thermal radiation reflected in the surface of the object in front of it. In a permanent installation, such as inside a device, the surface to be measured may have to be painted to reduce its reflectivity.

Glass Obstruction

Because glass is opaque to the spectrum of infrared that is of interest, temperature cannot be measured through a glass window. Silicon is opaque to visible wavelengths but is transparent to wavelengths longer than 2μ m.

Multiple Heat Sources

Heat is transferred by convection, conduction, and radiation. While an infrared temperature sensor is designed to be sensitive to radiation, it will also respond to other heat sources. Warm or cold air currents will affect its response, and so will heat conducted through the material on which it is mounted. Careful placement of the sensor is important. A shield around the sensor, with a small hole in the center, can prevent convection, while correct location on a circuit board can minimize conduction.

Thermal Gradients

An infrared temperature sensor should be mounted in a stable environment where it will not be exposed to thermal gradients (one side being hotter than another). This asymmetry can cause inaccurate readings.

microphone

29

OTHER RELATED COMPONENTS

- **speaker** (see Volume 2)
- **headphone** (see Volume 2)

What It Does

The sensation of sound is created by rapid waves of air pressure impinging upon the eardrum. A microphone can convert these pressure waves into an alternating electrical signal that can be amplified, recorded, broadcast, transmitted through wires, and reproduced as sound by a headphone or speaker. The principle is illustrated in [Figure 29-1](#). (For more information about sound reproduction, see the entries on **headphone** and **speaker** in Vol. 2.)

Schematic Symbol

Various schematic symbols for a microphone have been used during the decades since its invention. A selection is shown in [Figure 29-2](#). Each symbol assumes that sound is traveling from left to right. This is important when interpreting the symbol at top right, which can represent an earphone when it points in the opposite direction. Unfortunately, some schematics do not conform with this rule.

The two symbols at the bottom, showing a capacitor inside the microphone, should be reserved for condenser or electret microphones.

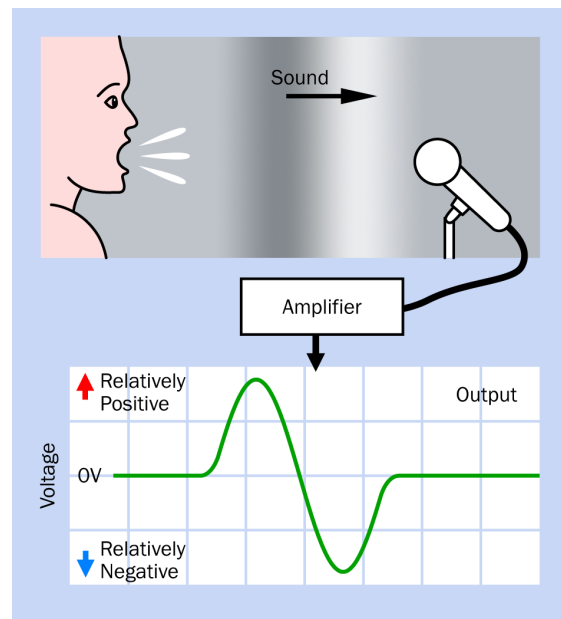


Figure 29-1 The principle of converting pressure waves into an alternating electrical signal (adapted from an illustration in [Make: More Electronics](#)).

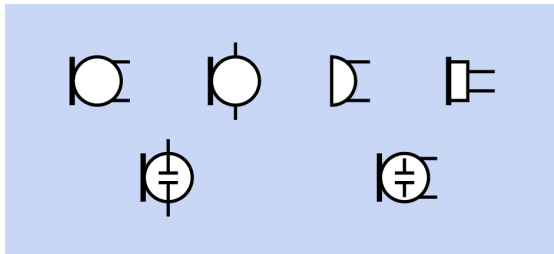


Figure 29-2 A selection of schematic symbols that represent a microphone.

How It Works

Some types of microphones generate a small voltage, while others have a fluctuating resistance that modulates a DC current.

Carbon Microphone

This was a very early attempt to reproduce sound. It contained carbon granules whose packing density increased and decreased in response to air pressure waves. When the density increased, the resistance between the granules diminished, and vice-versa. The principle is illustrated in [Figure 29-3](#), and was patented by Thomas Edison in 1877 for use in telephones. As late as the 1950s (and even later in some countries), wired telephone handsets contained carbon microphones. Their bandwidth was extremely limited.

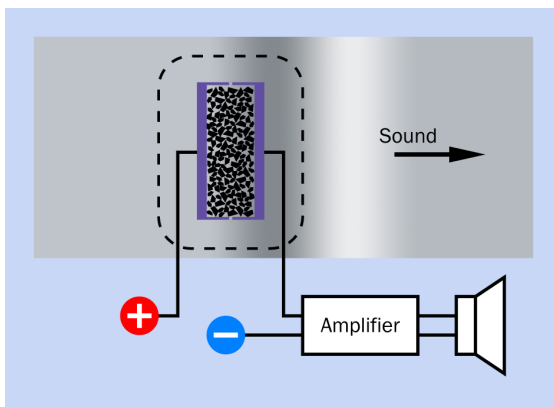


Figure 29-3 The principle of a carbon microphone.

Moving-Coil Microphone

Also known as a *dynamic* microphone, this consists of a very small, light coil of thin wire on a cylindrical tube that can vibrate along the axis of a permanent magnet. This principle is illustrated in [Figure 29-4](#). A diaphragm is attached to the front of the tube, and responds to air pressure waves that penetrate the perforated enclosure of the microphone. Movements of the coil around the magnet create small alternating currents in the wire. The inertia of the coil, tube, and diaphragm, and the force needed to overcome the interaction between the coil and the magnet, impose a limit on the high-frequency response of this design.

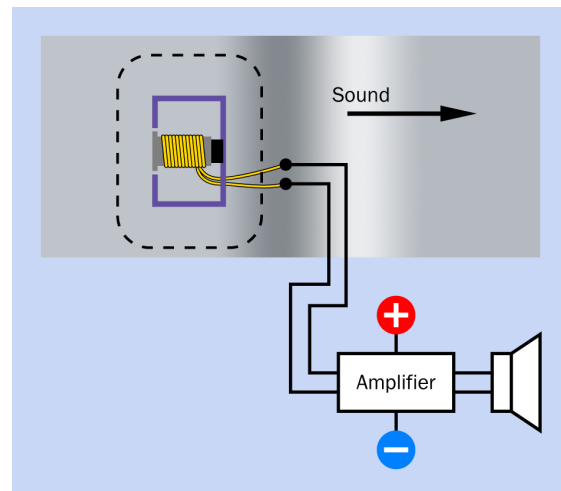


Figure 29-4 The principle of a moving-coil microphone.

Condenser Microphone

This type of microphone contains two thin discs or plates that form a capacitor. (In the early days of electricity, a capacitor used to be known as a condenser. The terminology has persisted for microphones.) An equal and opposite charge is applied to the plates. One plate is flexible, and as it responds to pressure waves, the capacitance between it and the other, rigid plate fluctuates. If the charge on the plates is kept approximately constant while the capacitance fluctuates, the voltage across the

capacitor fluctuates also. These fluctuations can be amplified, as suggested in [Figure 29-5](#).

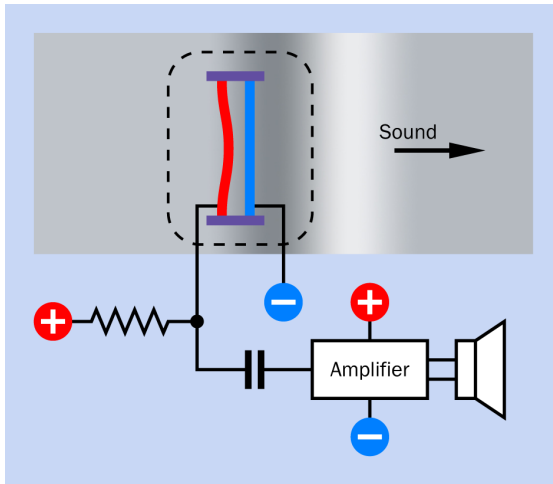


Figure 29-5 The principle of a condenser microphone.

Electret Microphone

This works on the same principle as a condenser microphone, except that its plates are made from a ferroelectric material that retains an electrical charge, just as iron will retain a magnetic polarization. The name of the microphone is derived from “electrostatic” and “magnet.” While early electret microphones were of poor quality, they have evolved to rival condenser microphones, and are extremely affordable. Because the electret creates very small currents, it usually includes a transistor or op-amp in its package to boost the signal, and has an open-collector output. The basic circuit for an electret is shown in [Figure 29-6](#). For more information about using an open-collector output, see [Figure A-4](#).

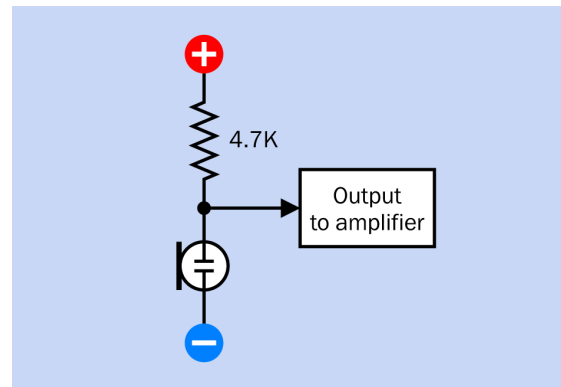


Figure 29-6 The basic circuit for using an electret microphone.

An example of a low-cost electret microphone is shown in [Figure 29-7](#). This type of component is sold either with leads attached, or with solder pads.

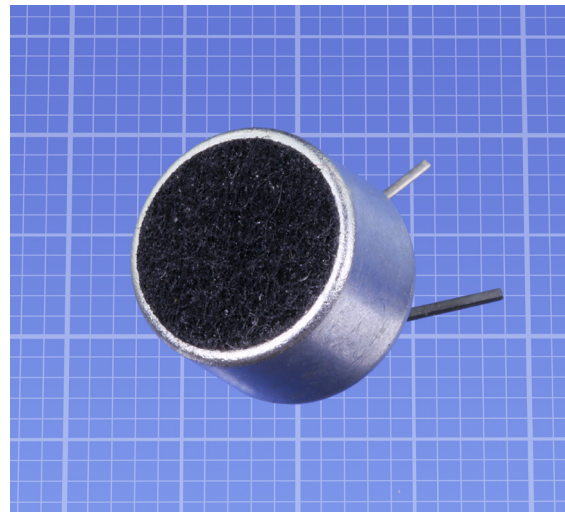


Figure 29-7 A generic electret microphone. The background grid is in millimeters.

MEMS Microphone

This type, often used in mobile phones, is a capacitive device that works on the same principle as a condenser microphone, although the component is etched in silicon and has a diaphragm that measures only about 1mm square. Many MEMS microphones have an analog output that is amplified in the same chip. Others have a digital output, using [PDM encoding](#). This

reduces the analog signal to a very fast bit stream, in which the density of the bits represents the amplitude of each fluctuation in a sound wave. PDM is an acronym for *pulse density modulation*. It requires an external clock signal to time the bit stream.

A breakout board from Sparkfun, on which is mounted an Analog Devices ADMP401 MEMS microphone with a preamplifier, is shown in Figure 29-8.

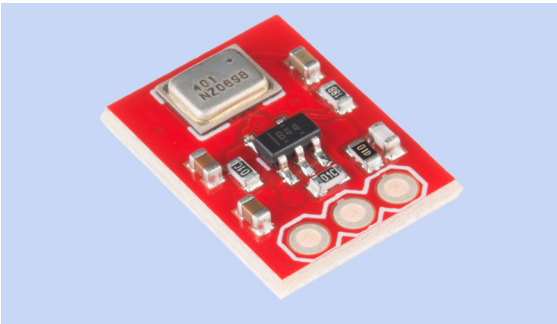


Figure 29-8 A breakout board for a MEMS microphone (the metal-clad rectangular component at the far end).

Piezoelectric Microphone

This has also been known as a *crystal* microphone. It contains a diaphragm that functions as a transducer. When it flexes in response to pressure waves, the mechanical energy is transformed into a small amount of electrical energy. Piezoelectric microphones were replaced by the moving-coil type in domestic audio devices when vacuum tubes were replaced with transistors, but may still be used as contact microphones to amplify acoustic musical instruments, or to trigger the playback of digitally sampled musical sounds.

Other variants include *ribbon* microphones (which were common in recording studios in the 1950s and 1960s, but have become rare), *laser* microphones, and *fiber-optic* microphones. They are not sufficiently common to be included in this entry.

Values

Sensitivity

Sound pressure is a complicated topic, explored in detail in the **transducer** entry in Volume 2. It can be measured in pascals, where 1 pascal = 1 newton per square meter.

The *sound pressure level* is a different concept. It measures the *relative* intensity of a sound, in a logarithmic scale calibrated in *decibels* (abbreviated dB). The reference value for this relative scale is 20 micropascals, considered to be the threshold of human hearing, comparable to a mosquito three meters away. This is assigned the value of 0dB.

From this point upward, the actual sound pressure doubles for each additional 6dB. A table of noise sources and their approximate decibel values is shown in Figure 29-9. This is derived from averaging eight similar tables, which are not always consistent in their estimates. It is an approximate guide only.

Decibels	Noise Example
140	Jet engine at 50 meters
130	Threshold of pain
120	Loud rock concert
110	Automobile horn at 1 meter
100	Jackhammer at 1 meter
90	Propeller plane 300 meters above
80	Freight train at 15 meters
70	Vacuum cleaner
60	Business office
50	Conversation
40	Library
30	Quiet bedroom
20	Leaves rustling
10	Calm breathing at 1 meter
0	Auditory threshold

Figure 29-9 Decibel values for some common sound sources. Reproduced from Volume 2 of this Encyclopedia.

The decibel unit is important when understanding the specifications of microphones, because it is used to measure their response. Microphone sensitivity is established with a standard input sine wave of 1kHz in frequency and 94dB in intensity (equivalent to 1 pascal in actual sound pressure), measured at the microphone. The sensitivity of an analog microphone is then defined as the number of decibels in an output signal of 1V. Because the output is an AC signal, voltage is measured as a root-mean-square (RMS) value.

For digital microphones, sensitivity is measured as the decibels that can be reproduced by a full-scale digital output. This value is abbreviated as dBFS.

Directionality

A microphone that has a directional response is desirable in many situations. Often, for example, sounds in front of the microphone are more important than sounds from behind the microphone. The directionality of a microphone (sometimes referred to as its *directivity*) is usually represented with a *polar graph* in which the microphone is seen from above, and its sensitivity to sounds from various directions is shown with a curve such as those in [Figure 29-10](#). The circles are drawn at intervals of 5dB, with 0dB at the periphery and -30dB at the center. The precise response for an individual microphone should be shown in its documentation.

Frequency Response

Every microphone tends to be more sensitive to some sound frequencies than others. A manufacturer will provide a graph showing this sensitivity, in decibels, plotted against sound frequency on a logarithmic horizontal axis. Theoretically, human hearing extends from around 20Hz to 20kHz, but few people are actually capable of hearing the high end of that range, and 15kHz may be a more realistic limit for a young person, diminishing to 10kHz with middle age.

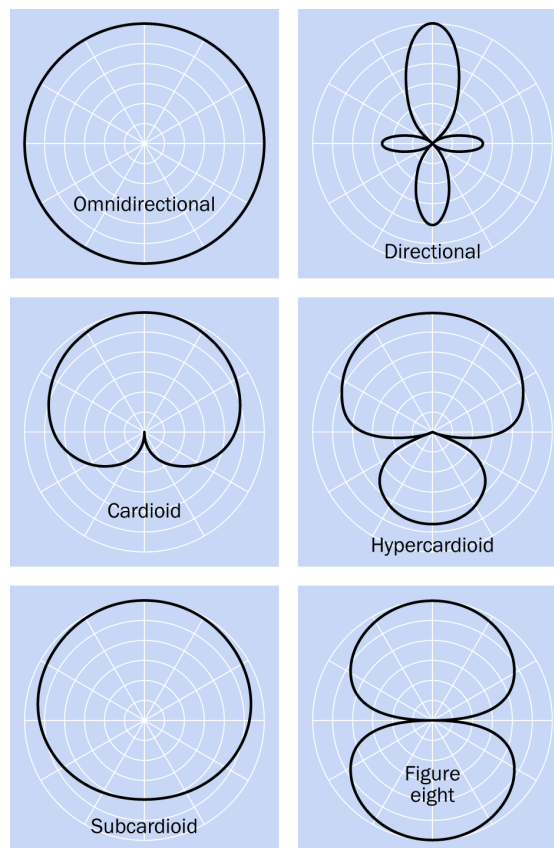


Figure 29-10 Six generic sensitivity patterns. Individual microphones will deviate somewhat from these generic curves.

An ideal flat response would show that a microphone is equally sensitive to all frequencies. In reality, *rolloff* usually occurs at low frequencies, and will eventually occur at high frequencies, although it may be preceded by a rise in response. If the central section of a curve is flat within plus-or-minus 1dB, this is a level of performance that was attained only by expensive studio microphones in the past. Electrets and MEMS microphones can now provide equivalent frequency response for \$1 or \$2 apiece, as opposed to the hundreds or even thousands of dollars that used to be necessary for professional equipment.

The response curve shown in [Figure 29-11](#) is for the eMerging i436, an electret microphone sold in a module as an accessory to enable high-

quality recordings on handheld devices. The rise around 15kHz may have been introduced deliberately by the manufacturer to compensate for reduced sensitivity of the human ear in that range.

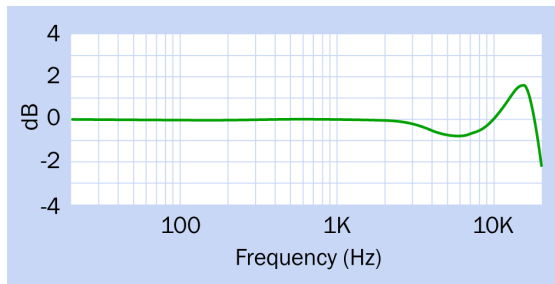


Figure 29-11 Frequency response for an electret microphone.

Impedance

The impedance value for a microphone is a function of its resistance, capacitance, and inductance. An amplifier input will also have an impedance rating, and for ideal power transfer between microphone and amplifier, the impedance values should be identical. However, a more important consideration in audio equipment is to avoid voltage loss between the output device (in this case, the microphone) and the input device (the amplifier). To achieve this, the output device should have a low impedance, while the input device should have a high impedance. Most microphones are rated at 150 to 200 ohms, while an amplifier may be rated at 1.5K to 3K.

Total Harmonic Distortion

When an audible sine wave is converted to an electrical output by a microphone (and by its preamplifier, if one is included in the module), the output may become corrupted by some multiples of the basic frequency. These are known as *harmonics*, and they are considered as a distortion of the signal. Total harmonic distortion, measured by a spectrum analyzer over the entire frequency range, should ideally be less than 0.01%.

Signal-to-Noise Ratio

Often abbreviated as S/N or SNR, the signal-to-noise ratio in a microphone is measured in decibels, and should be 60dB or higher.

What Can Go Wrong

Cable Sensitivity

Audio amplification is always vulnerable to electrical noise, which tends to be amplified along with the signal. Small signals from microphones require the use of shielded cables to reduce hum and other types of interference.

Noisy Power Supply

For similar reasons, power supplies must be as free as possible from voltage spikes and other fluctuations.

current sensor

30

The entry describes components that can be installed to monitor current on an indefinite basis. It does not include test equipment, such as test meters or multimeters.

A *current transformer* may be used to measure current, but is not included in this Encyclopedia.

OTHER RELATED COMPONENTS

- **voltage** sensor (see [Chapter 31](#))

What It Does

A current sensor measures the flow of electricity through a wire or a device, and supplies an output that can be interpreted either visually or by a microcontroller to provide a reading in amperes or fractions of an ampere.

Applications

Current sensing is important in industrial applications such as the control of high-powered motors. It can be used to monitor the performance of an inverter, or for everyday purposes such as monitoring the long-term power consumption of an appliance. During product development, a current sensor can indicate the power consumption of a circuit as it changes with modifications.

This entry describes three methods to measure current: an ammeter, series resistor, and Hall sensor. While other methods exist, they are outside the scope of this Encyclopedia.

Ammeter

An ammeter that is sold as a standalone device with leads for circuit testing is often described

as a *test meter*. Its functionality is usually built into a *multimeter*. Test meters and multimeters are outside the scope of this Encyclopedia.

An ammeter designed for permanent installation in a device or prototype is a type of *panel meter*, such as the one shown in [Figure 30-1](#). This traditional-style analog meter may be less expensive than the many digital types that are available. It uses a magnetic field created by current flowing through a coil to pull a needle across a scale, against the force of a spring.

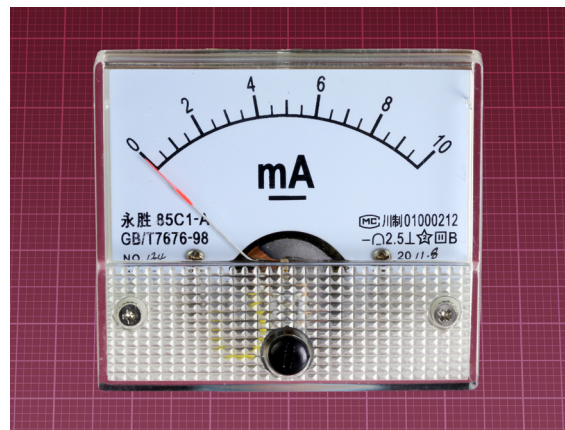


Figure 30-1 A traditional-style analog ammeter.

A digital ammeter allows a wider range of values to be viewed more easily. The meter from Adafruit in [Figure 30-2](#) has a range of 0A to 9.99A, at voltages from 4.5VDC to 30VDC. The meter can be powered parasitically from the currents that it is measuring, or can use a separate isolated 5VDC supply.

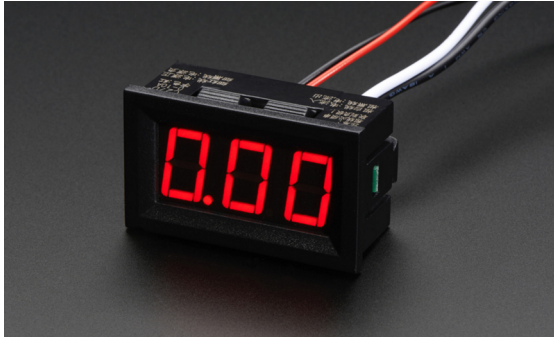


Figure 30-2 A panel-mount digital ammeter that can measure up to 9.99A.

Schematic Symbol

An ammeter may be represented in a schematic with the letter A inside a circle, as shown in [Figure 30-3](#).

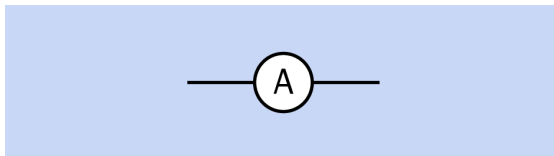


Figure 30-3 An ammeter may be represented like this in a schematic.

Ammeter Wiring

Two ways to use an ammeter in a circuit are illustrated in [Figure 30-4](#), where the load may be any equipment, device, or component that provides some electrical resistance. Because current is the same at all points in a simple circuit, the current that the meter measures flowing through itself will be the same as the current flowing through the load, and the sequence of components is immaterial.

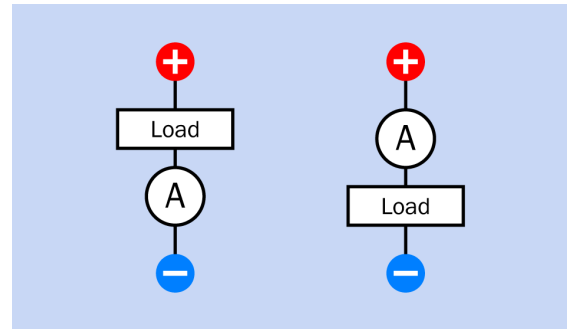


Figure 30-4 Two options for placement of an ammeter in a circuit.

However, regardless of the placement of the meter, the process of measuring current will inevitably change the value of the current being measured. This is because the ammeter imposes some internal resistance of its own. The resistance is extremely low, and may be considered negligible for loads of more than a few ohms.

- The low internal resistance of an ammeter means that it must never be connected in parallel with a load, or directly across a power source.

A disadvantage shared by analog and digital meters is that they are not usually interchangeable between AC and DC.

Series Resistor

The current flowing through a load can be calculated by measuring the voltage across a series resistor that is inserted between the load and its ground connection. The concept is illustrated in [Figure 30-5](#).

Using Ohm's law, if U is the voltage drop, I is the current, and R is the value of the resistor:

$$I = U / R$$

This tells us that for a fixed value of R, current is proportional to voltage. Therefore, measuring the voltage enables calculation of the current, so long as the value of R is known.

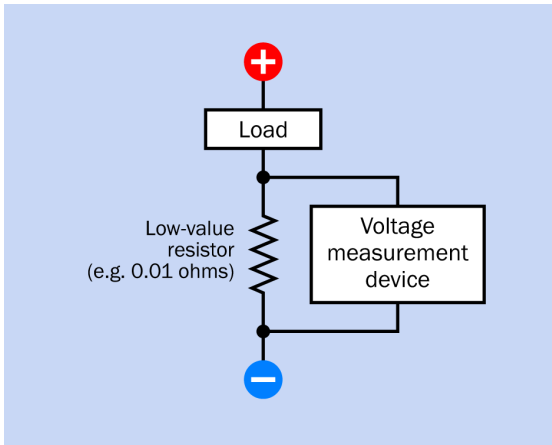


Figure 30-5 A basic circuit for measuring current as a function of the voltage drop across a small-value resistor. The load in this figure is any circuit or device with a relatively higher resistance. The voltage measurement device could be a microcontroller or analog-to-digital converter.

Suppose that R has a very small value, which is trivial compared with the resistance of the load. Consequently, the current in [Figure 30-5](#) will be mainly determined by the load, and we may consider the value of the current to be almost the same with or without the addition of R . In that case, the voltage drop across the resistor will be smaller if R is smaller. A smaller voltage drop will not be as easy to measure, but a lower resistance will result in less wastage of power.

If P is the power:

$$P = R * I^2$$

An example may help to make this clear. Suppose the resistor has a value of 0.5 ohms, and the voltage drop across it is measured to be 1V. Ohm's law shows that current flow is $1 / 0.5 = 2A$. The power formula shows that $P = 0.5 * 4 = 2W$.

To waste less power, the value of the resistor should be reduced further. Suppose a resistor of 0.01 ohms is used, and the voltage drop across it is measured as 0.02 volts. The current is $0.02 / 0.01 = 2A$, as in the preceding example, but the power dissipation is now only $0.01 * 4 = 0.04W$, which is negligible.

But are resistors available that have values measured in fractions of an ohm?

Current-Sense Resistors

In fact, many *current sense* resistors are available, with values of 0.1 ohms, 0.001 ohms, 0.0001 ohms, and many in between. Some resistors have values measured in micro-ohms. Examples are shown in [Figures 30-6](#), [30-7](#), and [30-8](#).

Measuring a small voltage drop is easily done by using a microcontroller. However, the connection to the microcontroller must be made as near to the resistor as possible, to eliminate the additional resistance of wires or circuit-board traces. For this reason, precision current-sense resistors may be equipped with four terminals. Two are wider, and are meant for connection to the flow of current. The other two are narrow, for measuring the voltage over the resistor. With this *4-point* configuration, voltage drop over the resistor can be measured as close to it as possible. The 0.001-ohm surface-mount resistor in [Figure 30-8](#) is designed for 4-point measurement.



Figure 30-6 Two current resistors manufactured by KOA Speer. Left: 0.1 ohms, 5W, 5%. Right: 1 ohm, 5W, 5%. The background grid is in millimeters.



Figure 30-7 Two styles of current resistors rated for 0.01 ohms. Bottom left: A plug-in version from TT Electronics, rated 1W and 5%. Top right: Ohmite, 4W, 1%. The background grid is in millimeters.

Resistors that have the lowest values and are intended to tolerate high current may consist of just a metallic strip welded to solderable pins. This type of component is sometimes called an *open-air resistor*. It is commonly used in multimeters, for measuring currents up to 10A or above.

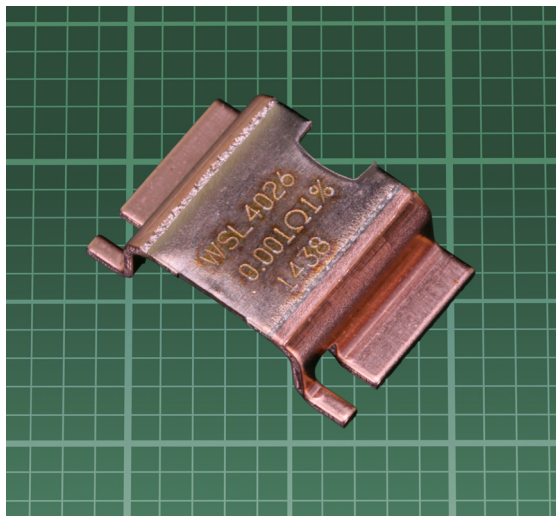


Figure 30-8 This Vishay 4-terminal surface-mount current resistor is rated for 0.001 ohms, 3W, 1%. The background grid is in millimeters.

Voltage Measurement

Some chips are designed for amplifying the voltage drop across a current-sensing resistor. An example is the Texas Instruments INA169.

A few chips contain an analog-to-digital converter in addition to the amplifier. The INA219 by Texas Instruments is designed to measure voltage as well as current, on the “high side” of a circuit—that is, between the positive power supply and the power input of the circuit. It makes its digital data available over an I2C bus.

For additional details about protocols such as I2C, see [Appendix A](#).

Measuring current from the voltage drop across a series resistor offers the advantages of simplicity, ability to work with AC or DC, and low cost (although some resistors of extremely low value can be relatively expensive). A possible disadvantage is that the measurement circuit is not isolated from the circuit whose current is being measured.

Hall-Effect Current Sensing

The principle of a Hall-effect sensor is explained in the entry for **object presence** sensors. See “[Hall-Effect Sensor](#)”. Normally this type of sensor is activated by an external permanent magnet, but it can also react to the magnetic field generated by current flowing through a wire.

Because the field generated around a wire is proportional with the current, the analog output voltage generated by a linear Hall-effect sensor can also be proportional with the current.

Hall-effect sensors for this specific purpose are available in 8-pad surface-mount packaging. The current to be measured passes through a copper conductor that is embedded in the chip. An example is Allegro’s ACS712, for AC or DC currents up to 30A. The internal resistance of the current path through this chip is stated

as 1.2 milliohms, and the path is isolated from the sensing circuitry.

Three variants of the chip are available, for currents of plus-or-minus 5A, 20A, and 30A. Depending on which variant is used, the output will range from 66mV to 185mV for each increase of 1A in the current path. Because the current path is isolated, the chip requires a separate power supply of 5VDC.

The 5A version of the AVS712 can be bought on breakout boards from Sparkfun. In [Figure 30-9](#), the board at top-left contains only the ACS712, while the board at lower-right adds an op-amp with sensitivity control, to amplify the voltage output when measuring small currents.

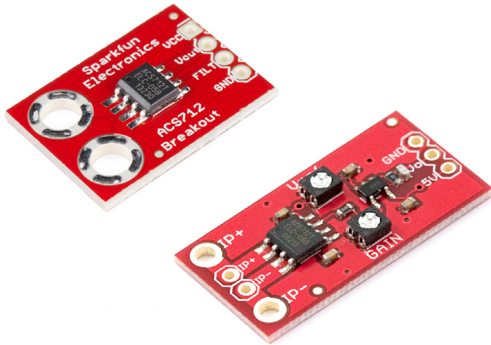


Figure 30-9 Breakout boards using the AVS712 Hall-effect current sensor. The one at lower-right includes an op-amp to amplify small signals.

What Can Go Wrong

Confusing AC with DC

A panel meter that is designed only to measure DC should not be used with AC, and vice-versa.

Erroneous readings or damage to the meter may result.

Magnetic Interference

A disadvantage of Hall-effect current sensing is that the sensor can be affected by stray magnetic fields. Because a Hall-effect chip is responding to very small magnetic effects, it is vulnerable to interference. Consult the manufacturer's datasheet carefully regarding correct placement of a chip on a circuit board.

Incorrect Meter Wiring

The correct wiring of an ammeter is in series with a load, not in parallel with a load. This may seem an elementary error, but is easy to make if an ammeter and a volt meter are both being used, and they look similar.

Because some panel meters are not fused, applying an ammeter directly across a power supply, without any series resistance, may result in immediate and hazardous destruction of the meter.

Incorrect wiring can also occur when using a digital meter that has four wires emerging from it: two for current testing and two for a separate power supply. This issue is especially important when measuring substantial currents (above 1A).

Current Out of Range

Attempting to measure a current that exceeds the range of an ammeter may damage the meter or blow its internal fuse, if it has one.

voltage sensor

31

The entry describes the type of component that can be installed to monitor voltage on an indefinite basis. It does not include test equipment, such as test meters or multimeters.

OTHER RELATED COMPONENTS

- **current** sensor (see [Chapter 30](#))

What It Does

A voltage sensor measures the electrical potential between any two points in a circuit, or the voltage supplied by a power source, and provides data in volts or fractions of a volt. It should not be confused with an *analog-to-digital converter*, which is not a sensor in itself, but can process the voltage output from a sensor by digitizing it. More information about analog-to-digital conversion is in the Appendix. See [Appendix A](#).

Applications

Voltage measurement is important in conjunction with all types of power supplies, to verify their performance. A volt meter may also be used to show the output from various types of analog sensors that have voltage output.

A graphical display may be used in audio equipment to indicate signal level, which can be proportional with voltage.

Volt Meter

A volt meter that is sold as a standalone device with leads for circuit testing is often described as a *test meter*. Its functionality is usually built

into a *multimeter*. Test meters and multimeters are outside the scope of this Encyclopedia.

A volt meter designed for permanent installation in a device or prototype is a type of *panel meter*, which is described here.

An antique analog panel meter is shown in [Figure 31-1](#).



Figure 31-1 An antique analog volt meter.

The four scales on the dial correspond with separate input terminals at the rear of the unit. The

unequally divided scales are a simple way to compensate for the nonlinear response of the mechanical movement inside the meter.

Modern analog volt meters are still manufactured, and may be cheaper than their digital equivalents. However, a digital volt meter allows a wider range of values to be viewed more easily. The meter in [Figure 31-2](#), sold as a low-cost battery tester, measures voltages from 4VDC to 13VDC to an accuracy of two decimal places. It needs no separate power supply.



Figure 31-2 A panel-mount digital volt meter.

Sometimes the number of digits in a volt meter is specified as ending with one-half, as in 3.5 or 3-1/2 digits. This means that the most significant (leftmost) digit can only be a 1 or a blank. The extra “half digit” may seem inconsequential but doubles the range of displayable values. For example, a 2-digit display can only show 100 values, from 0 to 99. A 2-1/2-digit display can show 200 values, from 0 to 199.

Schematic Symbol

A volt meter may be represented in a schematic with the letter V inside a circle, as shown in [Figure 31-3](#).

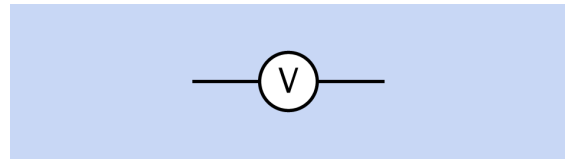


Figure 31-3 A volt meter may be represented like this in a schematic.

Volt Meter Wiring

Two ways to use a volt meter in a circuit are illustrated in [Figure 31-4](#), where the load may be any equipment, device, or component that provides some electrical resistance.

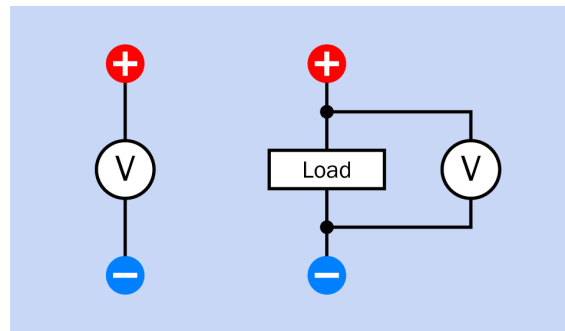


Figure 31-4 Two options for placement of a volt meter.

On the left side of the figure, a volt meter may be connected directly across a power source, because the internal resistance of the meter is so high, it will draw very little current. In this configuration, the meter measures the voltage of a source when the source is virtually unloaded.

On the right, the meter can measure the voltage drop across a load, or can measure the voltage across any component, or group of components, that are a subset of the load.

A disadvantage shared by dedicated analog and digital volt meters is that they are not usually interchangeable between AC and DC.

How It Works

An analog volt meter usually contains a built-in, high-value fixed resistance, and an ammeter

measuring the current passing through it. The current sensed by the ammeter is then converted to a reading in volts.

Suppose that R represents the fixed resistance, U is the voltage drop across the resistance, and I is the current flowing through it. Ohm's Law tells us:

$$U = I * R$$

The formula shows that if the resistance is fixed, the voltage will vary in proportion with the current, and therefore can be calculated from it.

Load-Related Inaccuracy

When the meter is measuring the voltage drop created by a load that has high resistance comparable to the internal resistance of the meter itself, the meter will not give an accurate reading. Figure 31-5 illustrates this problem.

On the left side of the figure, two resistors 10M each are wired in series between a 9VDC power source and negative ground. Because the resistors are equal, each of them imposes an equal voltage drop of 4.5V.

On the right side of the figure, an additional 10M resistance has been added in parallel with the lower resistor. When two resistors of values R_1 and R_2 are wired in parallel, their total resistance, R , is given by this formula:

$$1 / R = (1 / R_1) + (1 / R_2)$$

Therefore, the total resistance in the bottom half of the circuit is now 5M instead of 10M, and the voltage drop across the upper resistor becomes twice the voltage drop across the two lower resistors.

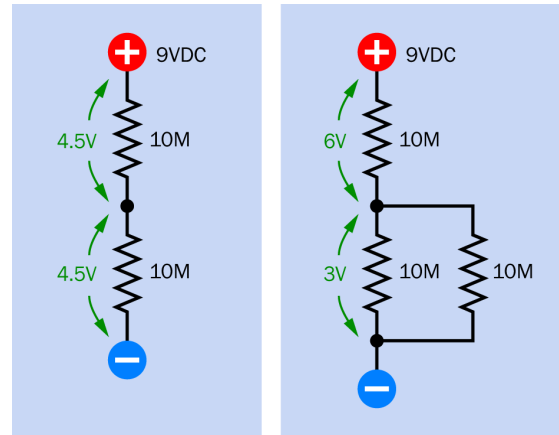


Figure 31-5 If a meter measuring the voltage drop across a load has an internal resistance comparable to that of a load, the meter will not give an accurate reading. See text for details.

Now suppose that the additional 10M resistor is actually the fixed resistance inside a volt meter. In fact, many volt meters do have an internal resistance of around 10M. Because the meter has reduced the resistance in the bottom half of the circuit by a factor of 2, it measures the voltage drop as 3V. If the meter had an ideal, infinite resistance, it would give the correct value of 4.5V. In the real world, that is impossible.

Bar Graph

Sometimes it is useful to represent voltage with a graphical display. A *bar graph* component makes this possible, and is often used in audio equipment.

The bar graph can consist of a row of *LEDs*. In a component designed for this purpose, there are usually 10 or more. To represent 0 volts, all the LEDs remain dark. More LEDs are illuminated as the voltage increases.

Because the bar graph itself only contains LEDs, a driver is necessary to convert voltage into the “rising thermometer” effect. Examples are the LM3914 (linear), the LM3915 (logarithmic, 3 dB per step), and the LM3916 (using VU or Volume Units, for audio). Each of these components has 10 LED outputs, and is fabricated using a chain

of resistors as a multitap voltage divider, with 10 comparators.

A driver can be set to show only one LED corresponding with the current voltage level, or (more commonly) a cumulative string of LEDs from zero upward. Some bar graphs also contain different colored LEDs, such as green for the first 7, yellow for the next 2, and red for the final one.

A microcontroller containing an analog-to-digital converter can be used to illuminate a bar graph, instead of a driver chip.

An example of a bar graph is the Avago HDSP-4830, shown in [Figure 31-6](#).

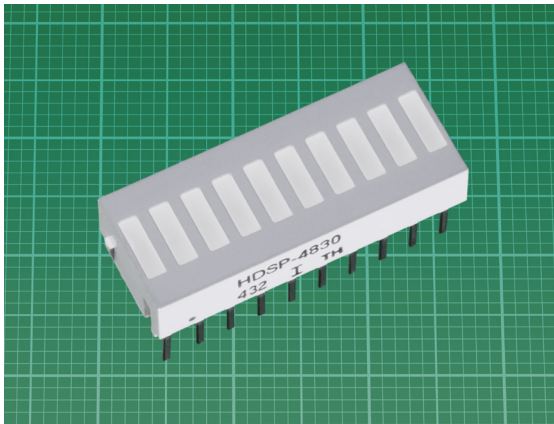


Figure 31-6 A bar graph LED display, which can be used to represent voltage. The background grid is in millimeters.

What Can Go Wrong

Confusing AC with DC

A panel meter that is designed only to measure DC should not be used with AC, and vice-versa. Erroneous readings or damage to the meter may result.

High Circuit Impedance

When a meter is measuring the voltage drop across a high-value resistance, it will give an inaccurate reading, as it is diverting a significant proportion of the current.

Voltage Out of Range

Attempting to measure a voltage that exceeds the limited range of a meter can damage the meter or blow its internal fuse, if it has one. Attempting to measure a voltage that is at the very low end of a meter's range will result in an inaccurate value.

Voltage Relative to Ground

If one input to a digital meter is connected with ground, the meter will only be able to measure voltage in a circuit relative to ground.

Sensor Output



This appendix provides some basic information about nine forms of sensor outputs, and the ways in which they can be processed. Other types of encoded output exist, but those examined here are the ones most likely to be found.

Figure A-1 provides an overview. Every sensor initially creates an analog output, which is sometimes connected directly to an output pin. In thermistors and photoresistors, for example, the internal resistance of the component constitutes its output. In many sensors, however, the behavior of the sensing element is processed internally to create voltage, open collector, encoded pulse stream, or current output.

If the sensor is chip-based, it may process the analog sensor response internally to create a binary output or a digital output.

In this Encyclopedia, the term “binary output” means “an output that has two states,” usually logic-low or logic-high. The states may be accessible via an output pin, or may be processed internally to create a pulse stream. In that case the stream fluctuates between the two states as a way of encoding an analog value with pulse-width modulation (PWM) or frequency. Other types of encoding are also possible, but are unusual.

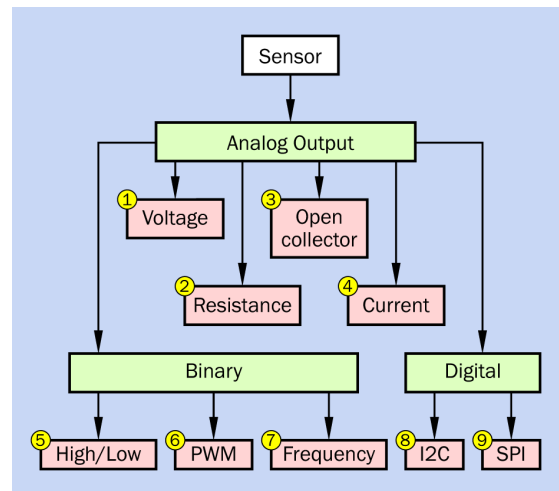


Figure A-1 Nine possible types of output from sensors. The primary categories are highlighted in green.

The term *digital output* is used here to mean one or two bytes of data that are stored in a register (memory location) in the sensor chip. While other forms of output are “always on” and can be accessed at any time from an output pin, a digital output is usually not available until an external device, such as a microcontroller, sends an instruction to the sensor chip, telling it to return the data. This two-way communication usually is handled by the I2C or SPI protocols (other protocols exist, but are less common).

Analog Outputs

Voltage is by far the most common form of analog output. Other forms of analog output can easily be converted to a voltage value, using the simple techniques described here.

1. Analog: Voltage

Direct Connection: Analog-to-Analog

An analog voltage output can be connected directly to an analog input, so long as the range is compatible and the sensor can provide sufficient current. Examples of external analog devices would be an analog volt meter, a light source or sound source that changes intensity, or a transistor or op-amp that will amplify the output for other audio/visual purposes.

If the voltage output from the sensor rises above a usable range, it can be converted to a lower value by applying it across two resistors connected in series to form a voltage divider. This is shown in [Figure A-2](#).

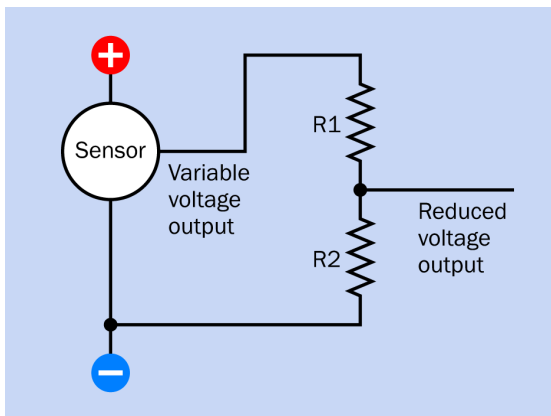


Figure A-2 Using a voltage divider to reduce the range from a sensor.

The values of R1 and R2 can be derived from this basic formula, where V_{SEN} is the voltage from the sensor and V_{OUT} is the output from the voltage divider:

Example A-1

$$V_{OUT} = V_{SEN} * (R2 / (R1 + R2))$$

The impedance of the device used for voltage sensing must be high relative to the values of R1 and R2. Note also that if the analog voltage out from the sensor varies linearly with the phenomenon being sensed, this relationship is likely to be disturbed by the voltage divider.

Analog-to-Binary Conversion

The term “binary” is used here to mean an output that can be in one of two states, such as logic-high and logic-low.

A varying analog voltage output can be simplified by passing it through a component that transforms the signal into binary form. This may be done by using a logic chip with a *Schmitt trigger* input, a *zener diode*, or a **comparator**. (For a description of comparators, see the entry in Volume 2.) A comparator provides desirable features such as adjustable positive feedback to create hysteresis. It may be used, for example, to convert a slowly changing signal from a phototransistor while the sun is setting, to a high/low output that can activate a relay to switch on a light.

Analog-to-Digital Conversion

The analog voltage output from a sensor can be digitized by an external analog-to-digital converter (ADC), either inside a microcontroller or by a separate ADC chip.

If a microcontroller is used, a sensor may often be wired directly to an input pin that connects internally with an ADC. A program in the microcontroller can then assess the integer output from the ADC and either execute a conditional statement or convert the value to a format appropriate for another device, such as a digital display.

If an ADC chip is used, there are thousands to choose from. A few basics:

- A *flash* converter contains a row of comparators with different reference voltages generated with a chain of equal resistors. The comparator outputs are fed into a priority **encoder**

that outputs a binary number. This system is very fast but has limited resolution.

- A *successive approximation* converter uses a single comparator, comparing the input voltage with the output from a DAC. The binary number that is supplied to the DAC is determined one bit at a time, from the most significant to the least significant bit, using the comparator's result to determine if the bit should be 0 or 1. These bits are stored in a register, called a successive approximation register (SAR). When the process finishes, the SAR contains a binary representation of the input voltage. This type of ADC can achieve high resolution (many bits) at the cost of lower conversion speed.
- In a *dual slope* converter, a capacitor is charged for a fixed time, with a rate proportional to the input voltage, then discharged at a known rate while measuring time by counting clock pulses. The resulting count is the ADC output.

The name of the converter is derived from the voltage on the capacitor.

- A *voltage-to-frequency* converter uses a voltage-controlled oscillator to produce pulses with a frequency proportional to the input voltage. If the pulses are counted over a fixed time interval, the count is proportional to the signal level.

The number of the bits in the output from an ADC must be sufficient to digitize the input voltage range with desired accuracy. Because the voltage range may contain unexpected peaks, a cautious strategy is to use many more bits than are necessary. However, this means that for most of the time, only a few bits will be used to represent the low end of the voltage range, and accuracy will suffer.

For example, suppose a voltage input normally ranges from 0V to 2V, with occasional brief excursions to 8V. An 8-bit ADC can provide 256 digital values to represent input voltages. If the values are spread uniformly over the full 8V input range, the least significant bit can measure 1/32nd of a volt, or about 31mV. Smaller voltage fluctuations will be ignored. On the other hand, if the 256 values are used to measure a range of just 2 volts, the least significant bit can measure 1/128th of a volt, or slightly less than 8mV—but voltages higher than 2V will be clipped.

An ADC will typically require a reference voltage, and will digitize the range from 0V to that voltage. The reference voltage must be chosen with the issues of accuracy and range in mind.

A microcontroller may provide a feature in its language to perform automatic scaling of an analog input within limits set by a variable in program code. This is done by comparing the input with a selectable voltage level, such as the voltage of the power supply, an externally supplied voltage, or a fixed built in reference. While the ADC in the microcontroller normally digitizes values from 0V to the supply voltage for the chip, the conversion routine can instruct the microcontroller to use its full number of bits (often, 10) to digitize an input range from 0V to 1V.

For a higher sample rate, an ADC chip may be connected with the microcontroller over an I2C or SPI bus.

2. Analog: Resistance

Resistance-to-Voltage Conversion

A sensor that changes its resistance as it responds to its environment can be placed in a *voltage divider* to provide an analog voltage output. This is illustrated in [Figure A-3](#).

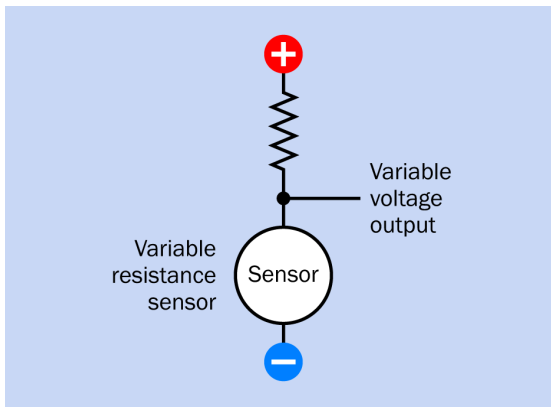


Figure A-3 The basic principle of putting a variable-resistance sensor in series with a fixed resistance, to create a voltage divider.

To choose the value for the series resistor, if R_{MIN} and R_{MAX} are the minimum and maximum resistance values for the sensor, the optimum value R_s for the series resistor, to produce the widest variation in voltages, will be found from this formula:

$$R_s = \sqrt{(R_{\text{min}} * R_{\text{max}})}$$

When the sensor has been set up in this way, the output can now be processed in the same way as any analog voltage output from a sensor.

The datasheet of the sensor should be consulted to make sure that a series resistor value does not allow any chance of the sensor being damaged by excessive current.

3. Analog: Open Collector

Many sensor packages or modules include a bipolar transistor that has an *open-collector* output (or *open drain*, if a CMOS transistor is used). The transistor may or may not be incorporated in an internal **op-amp** (described in Volume 2). Either way, the principle is the same.

Figure A-4 shows a sensor, in darker blue, that has one connection for positive power, another connection for negative ground, and a third

connection to the collector of the internal transistor.

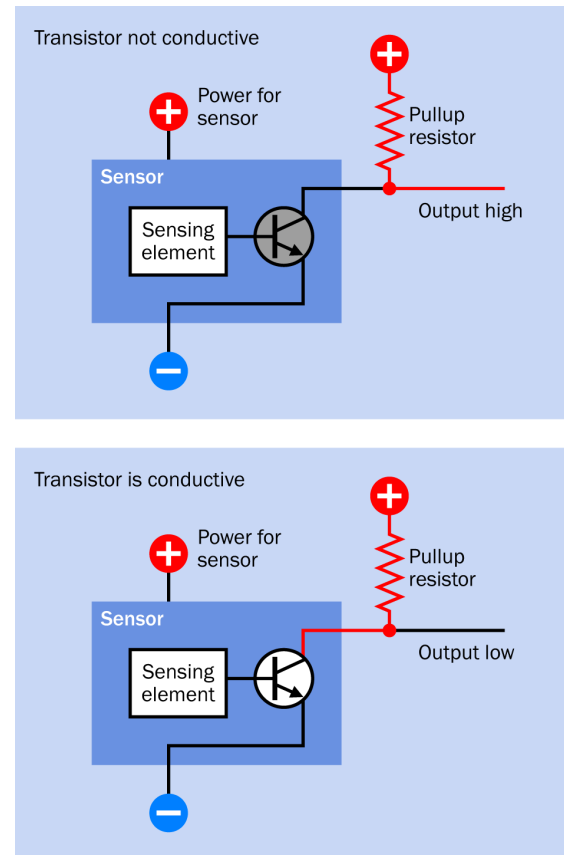


Figure A-4 How to use an open-collector output from an internal transistor.

In the upper section of the figure, the sensing element is not applying voltage to the base of the internal transistor, and the transistor only conducts a tiny amount of leakage current. Power applied to it through an external *pullup resistor* cannot reach negative ground in any significant quantity, and therefore it can provide a voltage input to a high-impedance device such as a microcontroller, or can power a component such as an LED, which draws relatively little current (20mA or less).

In the lower section of the figure, the sensing element is now applying voltage to the base of the transistor, drastically lowering its effective

resistance. The transistor diverts current from the pullup resistor to ground, and the output appears to go low.

The type of sensing element will determine whether the transistor becomes conductive or nonconductive when the element detects a stimulus.

The value that should be chosen for the pullup resistor will depend on the impedance of any device attached to the open-collector output. A 10K resistor may be appropriate for use with a device such as a microcontroller, which has very high impedance. At the other extreme, if the device is an LED, a 330-ohm resistor may be necessary. The value of the pullup resistor must be sufficiently low to enable reliable operation, but sufficiently high to prevent excessive current from passing through the internal transistor when it becomes conductive (20mA is a common maximum value).

The voltage from the open collector can be processed in the same way as any analog voltage output from a sensor.

An open-collector output may be used when the outputs from multiple devices share the same bus. One device can drive the bus without the problem of other devices attempting to hold the bus voltage high.

4. Analog: Current

Relatively few sensors provide an output consisting of variations in current. Some semiconductor temperature sensors function in this way. The output current can be converted to a voltage output simply by placing a fixed series resistor, as shown in [Figure A-5](#).

The voltage at the point shown, relative to negative ground, will vary linearly with the current. The value of the resistor should be defined in a datasheet for the sensor.

The voltage can now be processed in the same way as any analog voltage output from a sensor.

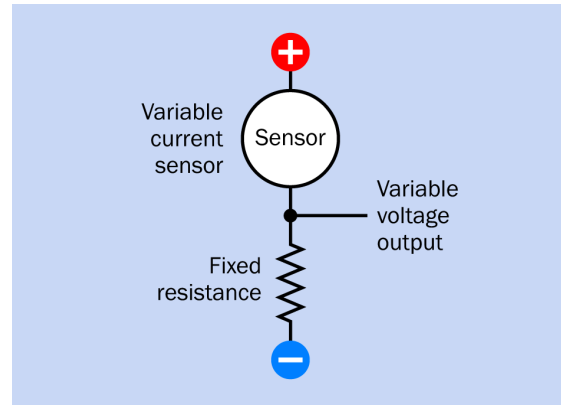


Figure A-5 How to convert the output from a variable-current sensor.

5. Binary: High/Low

A sensor that provides a binary output (that is, an output that is either logic-high or logic-low) can be connected directly with a microcontroller, if the voltage range is compatible. Program code in the microcontroller can then test the pin to establish its state. Note that some microcontrollers require a 3.3VDC power supply, while sensor chips may use 5VDC.

A binary output can also be used to control a solid-state relay, or an electromagnetic relay if transistor amplification is used. The output may be sufficient to power an LED indicator.

6. Binary: PWM

PWM is an abbreviation for *pulse-width modulation*. The sensor emits a stream of square-wave pulses with a fixed frequency, but the width of each pulse varies with the stimulus to which the sensor is responding. The width of each high pulse, relative to the wavelength between the start of one pulse and the start of the next, is called the *duty cycle*. A duty cycle of 0% means that there are no pulses at all. With a duty cycle of 100%, there are no gaps between pulses, so the output is high all the time. With a duty cycle of 50%, the duration of each high pulse is the same as the duration of the gap between pulses.

Various microcontrollers offer different ways to decode a PWM pulse stream. The most basic way is for a program to check an input pin repeatedly, as fast as possible, until a high state is detected. The microcontroller copies the value of its internal clock into a variable, then continues to check the input repeatedly until the pulse ends. The pulse duration that has been measured can be converted to a sensor value with a formula or lookup table.

This system is not recommended, as the microcontroller may miss the next pulse while it is converting the value of the previous pulse. To address this problem, a microcontroller language may offer a function that blocks execution of code while waiting for a pulse. The `pulseIn()` function on the Arduino is an example of this feature. However, the microprocessor must now spend most of its time waiting for a pulse instead of doing useful work.

A better solution is to write a program that is interrupt-driven.

Another option for decoding PWM is to use a low-pass filter that converts the pulse stream into an analog voltage, although some ripples will tend to remain.

Finally PWM can be used directly to power an LED or a DC motor, with transistor amplification as required. The speed of the motor or the brightness of the LED will vary with the duty cycle.

7. Binary: Frequency

Here again, the `pulseIn()` function on the Arduino may be used, so long as the frequency is a square wave with a known duty cycle.

8. Digital: I2C

In digital electronics, a *bus* is a communal pathway for sharing data among components or devices. The *I2C bus* is an abbreviation for *inter-integrated circuit* bus, developed originally by Philips in 1982. (Philips has since been subsumed into NXP Semiconductors.) The correct

notation for I2C is I²C, and it is spoken as “I-squared-C.” However, the term is very commonly written as I2C.

The I2C standard defines a data sharing protocol that is limited to 400kHz (with some exceptions) and designed to work on a small scale, almost always within one device, and usually on one circuit board. It is a low-cost, simple design. Data is transmitted serially over two wires, and the devices sharing the bus are connected in parallel.

Typically there is one *master* device on the bus, and a number of *slave* devices. Masters and slaves can both transmit information, but the master normally initiates communication. It also emits a clock signal for synchronization of data.

A sensor is a slave device that can be interrogated by a microcontroller in its role as the master device. Because multiple slave devices can share a bus, the microcontroller needs a way to identify the slave that it wishes to talk to, and each slave is assigned a unique address for this purpose. Often a slave will allow the user to modify the last two bits of the address, so that up to four identical devices can share a bus.

Code libraries to support the I2C protocol are available for most microcontrollers, and communication with a sensor that uses I2C should simply require knowledge of the sensor’s I2C address. However, the data registers in a sensor can be quite elaborate, requiring careful study of the manufacturer’s datasheet. Multiple procedures may be required to set functions on a device (such as the sensitivity range of an accelerometer, or the threshold for a temperature alarm). Multiple procedures may also be required to read data out of a sensor (such as two bytes to define a temperature, and several bytes to obtain time as well as location readings from a GPS module).

9. Digital: SPI

SPI is an acronym for serial peripherals interface, a standard introduced by Motorola that serves a similar function to the I2C bus, described immediately above. The SPI standard is slightly more sophisticated, enabling duplex communication and higher data transfer speeds. However, SPI requires a minimum of three wires shared by all devices on the bus, and an additional device-selection line for each slave device. The benefit of the extra device lines is that devices are easier to select and address than on an I2C bus, where more program statements are required. As is the case for

I2C, code libraries for microcontrollers are widely available to support SPI. However, the requirement for SPI to use three pins on a microcontroller, plus an additional pin for each slave device, is a disadvantage.

More sensors have I2C capability than SPI capability. The SPI protocol is potentially much faster than I2C.

A sensor that is SPI-enabled will almost certainly be available in a similar version that uses I2C. An increasing number of chip-based sensors support both protocols.

Glossary

This glossary is not comprehensive. It contains only the technical terms that have been used most frequently in this book in conjunction with sensor attributes.

ADC Analog-to-digital converter, which accepts a varying signal (usually, a voltage) as an input, and converts it to a digital value in the form of a binary number. This number is likely to have a high limit ranging from 255 decimal to 65535 decimal. Many microcontrollers contain their own ADC, which is multiplexed to assess the inputs on several pins. On an Arduino Uno, the ADC creates a digital value ranging from 0 through 1,023.

analog output When a sensor creates a voltage or varies its resistance without steps or increments, as a function of the phenomenon that it is measuring, this is an analog output.

binary output In this Encyclopedia, the term “binary output” describes a sensor output that only has two states: logic-high and logic-low, or on and off. The term is used in some data-sheets, but more often a binary output is described, misleadingly, as an analog output.

breakout board A small printed-circuit board containing one or more integrated circuit chips, usually surface-mounted. The board makes the features of the chips easier to access, because it has pins or connectors with 2.54mm (0.1”) spac-

ing for convenient experimental use with a breadboard. Additional features may be included, such as a voltage regulator.

chip-based sensor This term is used in the Encyclopedia to describe a sensor that is etched into a silicon chip and usually has signal conditioning components and circuitry built in.

contact bounce The tiny and rapid vibrations of mechanical switch contacts, when the switch opens or closes. If the switch is connected with a digital device such as a logic chip, debouncing hardware may be necessary, to allow the contacts time to settle. If the switch is connected with a microcontroller, a delay from 5ms to 50ms may be written into program code. Different switches have widely different settling times.

decibel A unit that expresses relative power or intensity, often (but not exclusively) applied to audible sound. The decibel is one-tenth of a bel, and is abbreviated dB, with the B capitalized because it is derived from the name of Alexander Graham Bell. Because dB is a logarithmic unit, the scale has no zero origin. However, 0dB may be assigned arbitrarily to any intensity, in which case lower intensities will have a negative value. An increase of 3dB corresponds to a doubling of sound intensity (acoustic energy). However, when a sound is sensed

by the human ear and evaluated by the brain, its subjective loudness doubles when the intensity increases by 10dB.

dielectric The insulating layer separating two plates in a capacitor.

hysteresis The difference between thresholds for switching an output on and off. When a sensor exhibits hysteresis, it may be unresponsive to a stimulus slightly above or below its equivalent current value. This may be useful to eliminate numerous responses to very small stimuli—for example, in a room thermostat.

I²C Interintegrated circuit bus. Sometimes written as I²C, and often referred to verbally as “I-squared-C.” A communications protocol that is often used between a microcontroller and other components on a circuit board. For a description, see “8. Digital: I²C” in the Appendix.

IMU Inertial measurement unit, consisting of three accelerometers and three gyroscopes, sometimes with the addition of three magnetometers. It can be used as a navigational aid. It may also be used in handheld user input devices, such as game controllers.

Kelvin A temperature scale, often abbreviated with the letter K, in which each degree is equivalent to a Celsius degree, but 0 degrees is at absolute zero—the temperature at which materials have no heat energy at all. 273 degrees K is approximately equal to 0 degrees C.

MEMS Microelectromechanical system, i.e., an integrated circuit chip that also contains tiny moving parts. For example, a MEMS accelerometer is built around microscopic springs that respond to accelerative forces.

newton A unit of force named after Isaac Newton, abbreviated with the capital letter N. A force of 1N will accelerate a mass of 1kg at a rate of 1 meter per second each second.

open collector output Many sensors have an open collector output, or contain an op-amp that has an open collector output. The output pin is attached to the collector of an internal transistor, with its emitter connected to negative ground. Positive voltage applied through a pullup resistor to the open collector will be grounded when the internal transistor is conducting current, but will be available for other devices when the transistor is off. See “3. Analog: Open Collector”.

orthogonal Angled at 90 degrees. Three orthogonal elements in a sensor will all be angled at 90 degrees to each other.

pascal A unit of pressure equivalent to 1 newton of force per square meter.

PIR Passive infrared sensor. See Chapter 4.

pullup resistor A resistor that pulls up an output or input voltage in the absence of a signal. May be used in conjunction with an **open collector output**.

quadrature An encoding system for output from a pair of sensors. If the sensors are identified as A and B, four output combinations are possible: A high and B low; A high and B high; A low and B high; A low and B low. A common application is to show the direction of movement of a magnetic or optical pattern past the sensor pair.

reference temperature The temperature at which the output signal of a temperature sensor is measured. This is often listed in a data-sheet.

register A section of memory that stores a digital value (usually 1 or 2 bytes in a sensor).

target Any object that is being detected by a motion sensor, proximity sensor, or presence sensor.

temperature coefficient The percentage increase or decrease in the value of a sensor as a result of unit change in temperature (usually 1 degree Celsius). Often abbreviated as TC. The

value may be resistance, voltage, or current, depending on the sensor. The temperature coefficient should be negative if the value of the sensor diminishes when its temperature increases. If it is expressed in parts per million, abbreviated ppm, it can be converted to a percentage by dividing by 10,000.

Wheatstone bridge A network of four resistors. At least one of the resistors has an unknown value, while the others have precisely known reference values. The network enables calculation of the unknown value. See [Figure 12-2](#).

Index

A

absolute humidity, 119, 120
absolute magnetic encoder, 41
absolute pressure, 113
accelerometer
 and gyroscope, 72
 applications, 70
 free fall, 71
 gravity, 71
 how it works, 70
 IMU, 5, 63, 66, 69
 Newton's Second Law, 72
 rotation, 71
 schematic symbol, 70
 spring, 70
 values, 73
 variants, 72
 what can go wrong, 74
 what it does, 69
ADC, 210, 217
air pressure, 114
alcohol vapor sensor (see gas concentration sensor, alcohol)
altimeter, 114
ammeter (see current sensor)
analog-to-digital conversion, 210, 217
anemometer, 125, 126
angle sensor (see rotary position sensor)
angular motion, 21

angular position sensor (see rotary position sensor)
Arduino-compatible
 gas flow rate sensor, 127
 gyroscope, 67
 IMU, 67
 magnetometer, 9
 pressure sensor, 114
 proximity sensor, 34
 temperature sensor, 184, 185
 touch screen, 97
 touch sensor, 91
avalanche diode, 136
axis of the earth, 7

B

bandgap temperature sensor (see semiconductor temperature sensor)
barometer, 114
barometric pressure, 114
Beidou positioning system, 3
bin switch, 57
black-body radiation, 25
Bourdon tube, 112
breakout board, 217
breakover switch (see tilt sensor)
Brokaw cell, 180

C

capacitive displacement sensor, 36
capacitive proximity sensor (see touch sensor)
capacitive touch screen, 96
capacitive touch sensor, 90
carbon microphone, 194
chip-based temperature sensor (see semiconductor temperature sensor)
clinometer, 55
coil magnetometer, 8
communications satellites, 1
compass, 6, 8
condenser microphone, 194
contact bounce, 18, 23
contents
 volume 1, xxi
 volume 2, xxii
 volume 3, xxii
crystal microphone, 196
current sensor
 ammeter, 199-200
 applications, 199
 Hall effect, 202
 how it works, 199
 microcontroller, 201
 panel meter, 199
 schematic, 200
 schematic symbol, 200

- series resistor, 200
- voltage drop, 200
- what can go wrong, 203
- what it does, 199

current-sense resistor, 201

D

debouncing, 18, 48, 61, 217

decibels, 196, 217

declination, 7

detection sensor (see presence sensor)

dew point, 119

dielectric, 90, 218

differential pressure, 113

diode temperature sensor (see semiconductor temperature sensor)

distance sensor (see proximity sensor)

dual slope converter, 211

dynamic pressure, 111

E

electret microphone, 195

encoder

- linear, 41
- magnetic, 41
- rotary, 48

errata, xxv

F

feedback to and from authors, xxv

flash converter, 211

float sensor, 100

float switch, 57

flux density, 6

flux, magnetic, 6

force sensor

- applications, 81
- deformative, 86
- how it works, 82
- how to use it, 86
- piezoelectric, 82

plastic film, 82, 85-87

resistive, 82, 86

schematic symbol, 82

strain gauge, 82, 83, 85, 88

units of measurement, 82

values, 87

what can go wrong, 88

what it does, 81

wheatstone bridge, 83

four-wire touch screen, 96

frequency response, 197

fresnel lens, 27

G

Galileo positioning system, 3

gas concentration sensor

- alcohol, 118
- breakout board, 119
- cross-sensitivity, 118
- how it works, 117
- humidity, 117, 119
- oxygen, 119
- propane, 118
- schematic symbol, 117
- semiconductor-based, 117
- what can go wrong, 122
- what it does, 117

gas flow rate sensor

- anemometer, 125, 126
- applications, 128
- data output, 128
- schematic symbol, 125
- thermal mass, 127
- thermopile, 127
- units of measurement, 128
- what can go wrong, 129
- what it does, 125

gas pressure sensor (see pressure sensor)

gas sensor (see gas concentration sensor)

gauge pressure, 113

geographical meridians, 7

Global Positioning System (see GPS)

GLONASS positioning system, 3

GPS

antenna, 2

definitions, 1

frequencies, 2

how to use it, 3

NMEA protocol, 3

satellites, 1

schematic symbol, 1

time-keeping, 4

tracker, 2

values, 3

what can go wrong, 4

what it does, 1

gray code, 51

gyroscope

- and accelerometer, 72
- applications, 63
- axes, 64
- how it works, 64
- how to use it, 67
- IMU, 5, 63, 66, 69
- resonator (see gyroscope, vibrating)
- schematic symbol, 63
- values, 67
- variants, 66
- vibrating, 64
- what can go wrong, 67
- what it does, 63

H

Hall effect

- current sensor, 202

Hall effect sensor

- advantages, 22
- applications, xx, 52
- float sensor, 100
- how it works, 18
- how to use, 20
- magnetometer, 8
- rotary position sensor, 48
- variants, 19

hard-iron bias, 9

heading, navigational, 5

humidity sensor (see gas concentration sensor, humidity)

hygrometer, 120

hysteresis

float sensor, 101, 218
reed switch, 17, 218

I

I²C, 214, 218
IC temperature sensor (see semi-conductor temperature sensor)
IMU, 5, 63, 66, 69, 218
inclination, 7
inclinometer, 55
inertial measurement unit (see IMU)
infrared proximity sensor, 32-34
infrared radiation, 188
infrared temperature sensor
 advantages, 187
 applications, 188
 array, 191
 compared with other temperature sensors, 153
 how it works, 188
 limitations, 188
 schematic symbol, 188
 thermopile, 189
 values, 191
 variants, 190
 what can go wrong, 192
 what it does, 187
inrush current limiter, 150, 161
interintegrated circuit bus (see I²C)

K

Kelvin temperature scale, 218

L

light meter, 137
light-dependent resistor (see photoresistor)
linear displacement sensor (see linear position sensor)
linear feedback shift register, 2
linear position sensor
 absolute, 41
 applications, 39, 42
 how it works, 39
 incremental, 41
 linear potentiometer, 40
 magnetic linear encoder, 41
 optical linear encoder, 41
 quadrature, 41, 218
 schematic symbols, 39
 what can go wrong, 43
 what it does, 39
linear potentiometer, 40
linear variable differential transformer, 42
liquid displacement sensor, 102
liquid flow rate sensor
 differential pressure, 109
 magnetic, 108
 paddlewheel, 105
 reed switch, 105
 rotor, 105
 schematic symbols, 105
 thermal mass, 107
 turbine, 106
 ultrasonic, 108
 what can go wrong, 109
 what it does, 105
liquid flow switch, 108
liquid level sensor
 displacement, 102
 float, 100
 float switch, 57
 how it works, 100
 incremental output, 101
 potentiometer, 101
 pressure of liquid, 103
 schematic symbols, 99
 tilting, 104
 turbulence, 104
 ultrasonic, 102
 weight of reservoir, 103
 what can go wrong, 104
 what it does, 99
liquid pressure sensor (see pressure sensor)
load cell (see load sensor)
load sensor, 81
 (see also force sensor)
 applications, 82
 deformative, 86

liquid level measurement, 103
schematic symbol, 82
strain gauge, 83, 85, 88
values, 87
what can go wrong, 88
what it does, 81

LVDT (see linear variable differential transformer)

M

magnetic
 axis, 7
 ball in tilt sensor, 59
 declination, 7
 encoder, 41, 52
 field, 6
 flux, 6
 liquid flow rate sensor, 108
 meridians, 7
 presence sensors, 16
 rotary position sensor, 47
magnetic, linear encoder, 41
magnetometer
 breakout board, 6, 9
 coil, 8
 hard-iron bias, 9
 heading, 5
 how it works, 6
 how to use it, 9
 IMU, 5, 63, 69
 magnetoresistance, 8
 meridians, 7
 placement, 10
 scalar, 5
 schematic symbol, 5
 soft-iron bias, 9
 vector, 5
 what can go wrong, 9
 what it does, 5
magnetoresistance, 8
mass flow rate sensor (see gas flow rate sensor)
mathematical syntax, xxiv
membrane switch, 89
MEMS, xix, 218
MEMS microphone, 195
mercury switch, 58, 61, 78

- meridians, 7
- metric system, xxiv
- microelectromechanical devices (see MEMS)
- microphone
 - carbon, 194
 - condenser, 194
 - crystal, 196
 - directionality, 197
 - dynamic, 194
 - electret, 195
 - frequency response, 197
 - how it works, 194
 - impedance, 198
 - MEMS, 195
 - moving coil, 194
 - piezoelectric, 196
 - rolloff, 197
 - schematic symbols, 193
 - sensitivity, 196
 - signal-to-noise ratio, 198
 - total harmonic distortion, 198
 - units of measurement, 196
 - what can go wrong, 198
 - what it does, 193
- motion sensor (see PIR)
- mouse, 50
- moving coil microphone, 194

N

- navigational heading, 5
- newton unit of force, 218
- Newton's Second Law, 72
- NMEA protocol, 3
- north pole, 7
- NTC thermistor (see thermistor, NTC)

O

- object detector (see presence sensor)
- object presence sensor (see presence sensor)
- open collector output, 218
- opt-pass sensor, 15
- optical

- linear encoder, 41
- optical isolator, 132
- optical presence sensor, 12
- optical rotary encoder, 48, 53
- optical switch, 13, 15, 50
- optocoupler, 142
- optointerrupter, 21
- orthogonal, 218
- output
 - analog current conversion, 213
 - analog voltage conversion, 210, 217
 - analog-to-binary conversion, 210
 - analog-to-digital conversion, 210
 - binary, 209, 210, 213, 217
 - digital, 209
 - dual slope converter, 211
 - flash converter, 211
 - frequency conversion, 214
 - I2C, 214, 218
 - open collector, 212, 218
 - options, 209
 - pullup resistor, 213, 218
 - PWM conversion, 209, 213
 - resistance-to-voltage conversion, 211
 - scaling, 211
 - SPI, 215
 - successive approximation converter, 211
 - voltage divider, 210
 - voltage-to-frequency converter, 211
- over-current protection, 160
- over-temperature protection, 159
- oxygen sensor (see gas concentration sensor, oxygen)

P

- paddlewheel flow rate sensor, 105
- panel meter, 199, 205
- pascal unit of pressure, 218
- passive infrared motion sensor (see PIR)
- pendulum switch, 58

- photocell (see photoresistor)
- photoconductive cell (see photoresistor)
- photodarlington, 142
- photodiode
 - applications, 135
 - array, 137
 - avalanche diode, 136
 - band gap, 137
 - color sensing, 137
 - dark current, 135
 - how it works, 135
 - how to use it, 139
 - light meter, 137
 - photoconductive mode, 135
 - phototransistor comparison, 143
 - photovoltaic mode, 135
 - PIN type, 136
 - reverse bias, 135
 - schematic symbols, 135
 - ultraviolet, 137
 - values, 138
 - variants, 136-137
 - wavelength range, 137
 - what can go wrong, 140
 - what it does, 135
- photointerrupter, 13
- photomicrosensor, 15
- photoresistor
 - availability, 131
 - how it works, 132
 - how to use it, 133
 - optical isolator, 132
 - phototransistor comparison, 133
 - schematic symbols, 131
 - sizes, 132
 - voltage divider, 134
 - wavelengths of sensitivity, 134
 - what can go wrong, 134
 - what it does, 131
- phototransistor
 - applications, 142
 - base connection, 142
 - binning, 144
 - FET, 142
 - how it works, 142
 - how to use it, 144

- optocoupler, 142
- output current calculation, 145
- photodarlington, 142
- photodiode comparison, 143
- photoresistor comparison, 133
- schematic symbols, 141
- values, 143
- what can go wrong, 145
- what it does, 141
- photovoltaic effect, 135
- piezoelectric
 - force sensor, 82
 - heat detector, 26
 - microphone, 196
 - ultrasonic transducer, 32
 - vibration sensor, 76
- PIN photodiode, 136
- pin-and-spring vibration sensor, 76
- PIR
 - applications, 25
 - fresnel lens, 27
 - how it works, 26
 - schematic symbols, 25
 - variants, 29
 - what can go wrong, 30
 - what it does, 25
- potentiometer
 - arc segment, 46
 - end stops, 46
 - float sensor, 101
 - logarithmic, 46
 - multiturn, 46
 - rotary position, 46
- presence sensor
 - angular motion, 21
 - configuration, 21
 - linear motion, 21
 - magnetic, 16
 - optical, 12, 22, 23
 - optointerrupter, 21
 - reflective, 12
 - retroreflective, 12, 15
 - schematic symbol, 12
 - transmissive, 12-13
 - what can go wrong, 23
 - what it does, 11
- pressure sensor
 - absolute pressure, 113

- altitude, 114
- ambient air, 114
- applications, 111
- barometric, 114
- Bourdon tube, 112
- diaphragm, 112
- differential pressure, 113
- dynamic, 111
- gas and liquid behavior, 112
- gases, 115
- gauge pressure, 113
- how it works, 112
- schematic symbols, 111
- sensing elements, 112
- static, 111
- units of measurement, 112
- what can go wrong, 116
- what it does, 111
- propane sensor (see gas concentration sensor, propane)
- proximity sensor
 - applications, 32
 - capacitive displacement, 36
 - infrared, 32-34
 - schematic symbols, 31
 - triangulation, 32
 - ultrasound, 32-33
 - what can go wrong, 38
 - what it does, 31
- PTC thermistor (see thermistor, PTC)
- pullup resistor, 213, 218
- pulse-width modulation (see PWM)
- PWM, 209, 213
- pyroelectric detector, 26

Q

- quadrature, 41, 49, 218

R

- range finder (see proximity sensor)
- reed switch
 - advantages, 22
 - alarm system, 17
 - float sensor, 100

- how it works, 17
- how to use it, 18
- hysteresis, 17
- liquid flow rate sensor, 105
- values, 18
- variants, 17
- what can go wrong, 23
- reference sources, xxvii
- reference temperature, 218
- reflective interrupter, 15
- reflective object sensor, 15
- reflective optical sensor, 15
- reflective photointerrupter, 15
- register, 218
- relative humidity, 119, 120
- resistance temperature detector (see RTD temperature sensor)
- resistance-to-voltage conversion, 211
- resistive force sensor, 82, 86
- resistive temperature device (see RTD temperature sensor)
- resistive touch screen, 95
- retroreflective presence sensor, 12, 15
- rotary encoder, 48, 52-53
- rotary position sensor
 - absolute, 51
 - accelerometer, 71
 - applications, 45
 - gray code, 51
 - Hall Effect sensor, 48
 - incremental, 49
 - magnetic, 47
 - mouse, 50
 - optical encoder, 48
 - potentiometer, 46
 - quadrature, 49, 218
 - reflective optical, 48
 - rotary encoder, 48
 - schematic symbols, 46
 - sensing chips, 48
 - transmissive optical, 48
 - what can go wrong, 53
 - what it does, 45
- rotary sensor (see rotary position sensor)
- rotating coil magnetometer, 8

- rotation sensor (see rotary position sensor)
- rotational encoder (see rotary encoder)
- rotational position sensor (see rotary position sensor)
- RTD temperature sensor
 - applications, 174
 - attributes, 173
 - compared with other temperature sensors, 153
 - how it works, 174
 - output conversion, 176
 - probe enclosure, 176
 - schematic symbol, 174
 - three-wire variant, 175
 - variants, 175
 - what can go wrong, 176
 - what it does, 173

S

- Safari Books Online, xxvi
- satellites, 1
- scalar magnetometer, 5
- schematic conventions, xxiv
- schematic symbol
 - accelerometer, 70
 - current sensor, 200
 - force sensor, 82
 - gas concentration sensor, 117
 - gas flow rate sensor, 125
 - GPS, 1
 - gyroscope, 63
 - infrared temperature sensor, 188
 - linear position sensor, 39
 - liquid flow rate sensor, 105
 - liquid level sensor, 99
 - load sensor, 82
 - magnetometer, 5
 - microphone, 193
 - photodiode, 135
 - photoresistor, 131
 - phototransistor, 141
 - PIR, 25
 - presence sensor, 12
 - pressure sensor, 111
 - proximity sensor, 31
 - rotary position sensor, 46
 - RTD temperature sensor, 174
 - semiconductor temperature sensor, 178
 - thermistor, NTC, 147
 - thermistor, PTC, 157
 - thermocouple, 166
 - tilt sensor, 56
 - touch screen, 95
 - touch sensor, 90
 - vibration sensor, 75
 - voltage sensor, 206
- Seebeck effect, 167
- semiconductor temperature sensor
 - analog current output, 182
 - analog voltage output, 181
 - applications, 178
 - attributes, 178
 - Brokaw cell, 180
 - CMOS type, 185
 - compared with other temperature sensors, 153
 - digital output, 183
 - how it works, 179
 - multiple transistors, 179
 - schematic symbol, 178
 - variants, 180
 - what can go wrong, 185
 - what it does, 177
- sensor
 - categories, xx
 - chip-based, 217
 - multifunction, xx
 - output options, 209
 - suppliers, xxv
- serial peripherals interface (see SPI)
- shift register, 2
- silistor, 158
- slider potentiometer, 40
- soft-iron bias, 9
- solar panel, 135
- sound pressure, 196
- sources, xxvii
- south pole, 7
- speed sensor, 45
- SPI, 215

- static pressure, 111
- stomp box, 133
- strain gauge, 82-83, 85, 88
- successive approximation converter, 211
- suppliers, xxv
- surge limiter (see inrush current limiter)

T

- tactile switch, 89
- target, 218
- temperature coefficient, 147, 152, 218
- temperature sensor
 - comparison of types, 153
 - infrared (see infrared temperature sensor)
 - noncontact type, 187
 - RTD (see RTD temperature sensor)
 - semiconductor (see semiconductor temperature sensor)
 - thermistor, NTC (see thermistor, NTC)
 - thermistor, PTC (see thermistor, PTC)
 - thermocouple (see thermocouple)
- thermal mass flow rate sensor (see gas flow rate sensor)
- thermal mass liquid flow rate sensor, 107
- thermistor, NTC
 - applications, 148
 - compared with other temperature sensors, 153
 - datasheet terminology, 152
 - how it works, 148
 - inrush current limiter, 150
 - output conversion, 149
 - schematic symbols, 147
 - values, 152
 - what can go wrong, 153
 - what it does, 147
 - wheatstone bridge, 150
- thermistor, PTC

- compared with other temperature sensors, 153
 - externally heated, 159
 - heating element, 162
 - inrush current limiter, 161
 - lighting ballast, 162
 - linear, 158
 - nonlinear, 158, 159
 - over-current protection, 160
 - over-temperature protection, 159
 - schematic symbols, 157
 - silistor, 158
 - starting current, 162
 - what can go wrong, 163
 - what it does, 157
- thermocouple
 - applications, 166
 - compared with other temperature sensors, 153
 - how it works, 167
 - how to use it, 168
 - identifiers, 169
 - measurement junction, 168
 - output conversion, 170
 - reference junction, 168
 - schematic symbol, 166
 - Seebeck coefficients, 169
 - Seebeck effect, 167
 - thermopile, 171
 - what can go wrong, 171
 - what it does, 165
- thermopile
 - gas flow rate sensor, 127
 - infrared temperature sensor, 189
 - overview, 171
- tilt sensor
 - applications, 57
 - bin switch, 57
 - breakout board, 59
 - float switch, 57
 - how it works, 56
 - how to use it, 61
 - magnetic ball, 59
 - mercury switch, 58, 61
 - miniaturized, 59
 - pendulum switch, 58
 - schematic symbol, 56
 - two-axis, 59
 - values, 60
 - what can go wrong, 61
 - what it does, 55
- tilt switch (see tilt sensor)
- tip sensor (see tilt sensor)
- tipover switch (see tilt sensor)
- touch pad (see touch sensor)
- touch screen
 - Arduino-compatible, 97
 - capacitive, 96
 - four-wire, 96
 - resistive, 95
 - schematic symbol, 95
 - variants, 95, 97
 - what it does, 95
- touch sensor
 - applications, 90
 - breakout boards, 91
 - capacitive, 90
 - design considerations, 92
 - how it works, 90
 - how to use it, 91
 - schematic symbols, 90
 - tactile feedback, 90
 - touch pad availability, 91
 - what can go wrong, 93
 - what it does, 89
 - wheels and strips, 92
- touch strip, 92
- touch wheel, 92
- triangulation, 32
- turbine flow rate sensor, 106

U

- ultrasonic anemometer, 126
- ultrasonic flow rate sensor, 108
- ultrasonic liquid level sensor, 102
- ultrasonic proximity sensor, 32-33

units of measurement, xxiv

V

- Vactrol, 132
- vapor sensor (see gas concentration sensor)
- vector magnetometer, 5
- vibrating gyroscope, 64
- vibration sensor
 - how to use it, 79
 - magnetic, 77
 - mercury switch, 78
 - mousetrap type, 77
 - piezoelectric, 76
 - pin-and-spring, 76
 - schematic symbols, 75
 - values, 78
 - variables, 78
 - variants, 75
 - what can go wrong, 79
 - what it does, 75
- vibration switch (see vibration sensor)
- video display, 95
- volt meter (see voltage sensor)
- voltage divider, 210
- voltage sensor
 - accuracy, 207
 - bar graph, 207
 - how it works, 206
 - panel meter, 205
 - schematic, 206
 - schematic symbol, 206
 - volt meter, 205
 - what can go wrong, 208
 - what it does, 205
- voltage-to-frequency converter, 211

W

- wheatstone bridge, 83, 150, 160, 219

About the Authors

Charles Platt is the author of *Make: Electronics* and *Make: More Electronics*. He is a former senior writer for *Wired* magazine, and is a contributing editor to *Make:* magazine, for which he writes a column on electronics.

Fredrik Jansson is a physicist from Finland, with a PhD from Åbo Akademi University. He is currently living in The Netherlands, where he works on swarm robotics and simulates sea animals in the Computational Science group at the University of Amsterdam. Fredrik has always loved scavenging discarded household electronics for parts, and is a somewhat inactive radio amateur with the call sign OH1HSN. He also fact-checked Charles Platt's previous book, *Make: More Electronics*.

Colophon

The cover and body fonts are Myriad Pro, the heading font is Benton Sans, and the code font is Dalton Maag's Ubuntu Mono.